SOCIAL STATISTICS

HARPER'S SOCIAL SCIENCE SERIES

F. STUART CHAPIN, EDITOR

HUMAN RELATIONS
by Carl C. Taylor
and B. F. Brown

RURAL SOCIOLOGY
(Revised Edition)
by Carl C. Taylor

AN INTRODUCTION TO ANTHROPOLOGY
by Wilson D. Wallis

SOCIOLOGY AND EDUCATION
by Alvin Good

SOCIAL MOBILITY
by Pitirim Sorokin

PROBLEMS OF SOCIAL WELL-BEING
by J. H. S. Bossard

CONTEMPORARY SOCIOLOGICAL THEORIES
by Pitirim Sorokin

SOCIAL WORK ADMINISTRATION
by Elwood Street

THE SOCIAL WORKER
IN FAMILY, MEDICAL AND PSYCHIATRIC SOCIAL WORK
by Louise C. Odencrantz

THE SOCIAL WORKER IN GROUP WORK
by Margaretta Williamson

TRENDS IN AMERICAN SOCIOLOGY
by George A. Lundberg and others

THE SOCIAL WORKER IN CHILD CARE AND PROTECTION
by Margaretta Williamson

AMERICAN MINORITY PEOPLES
by Donald Young

SOCIAL PSYCHOLOGY
by Joseph K. Folsom

PRINCIPLES OF SOCIOLOGY
by E. T. Hiller

SOCIAL STATISTICS
by R. Clyde White

SOCIAL STATISTICS

By

R. CLYDE WHITE

Professor of Sociology and Director of the Bureau of Social Research Indiana University

HARPER & BROTHERS PUBLISHERS

New York and London

1933

SOCIAL STATISTICS

Printed in the U. S. A.
First Edition

Editor's Introduction

STATISTICAL method has become a fundamental tool to scientific advances in sociology and in social work. This book is unique in that it combines between two covers divisions of statistical method which have hitherto been featured separately only in books on economic statistics, or in books on psychological statistics, or in books on vital statistics.

Professor White has woven into a single consistent treatment, not only the usual techniques of tabulation, graphic representation, the measurement of central tendencies, dispersion and correlation, but he has added a simple presentation of the technique of analysis of time series, an outline of the chief elements of vital statistics, and a suggestive treatment of the technique of social measurements and the standardization of sociometric scales.

F. STUART CHAPIN

Preface

A GREAT deal of work has been done during the last decade in the field of what may properly be called "social statistics" as distinguished from economic and business statistics.

Until recent years social statistics connoted only columns of figures. It still refers to the tabulation of social data, but it is an improved and extended tabulation plus a technique for extracting the meaning of such data, namely, the methods of statistical analysis applied to social data for scientific and practical purposes.

The forerunners of the present-day social statisticians were such men as Quételet, Pareto, Galton, Mayo-Smith, Giddings, and Wright. Giddings belongs in this list, not because he did a great amount of statistical work himself, but because of the influence he exercised in directing his students into quantitative studies of social data. Among the leaders in social statistics today may be mentioned Chaddock, Chapin, Dublin, Shelby Harrison, Hexter, Hurlin, Ogburn, Rice, Frank A. Ross, Dorothy S. Thomas, Truesdell, and many others. These men and their predecessors have created a special division of the field of statistics, and the result is that a large number of colleges and universities in the United States are now giving courses in this branch of statistics.

This book represents an effort to adapt statistical methods to the data of sociology and social work for teaching purposes in the light of the work done by American social statisticians. The methods and principles discussed and illustrated are well known. Whatever innovation there may be lies in the fact that ordinary methods of statistics have been applied systematically to the data of sociology and social work. The author believes that a social statistician learns his technique and acquires the correct habits of thought about his work by applying statistical methods to his own data. He is not likely to have much of a penchant for social statistics, if he learns statistical methods through the use of biological data. Familiarity with the data of sociology and social work and practice in analyzing these data by statistical methods are fundamental to the training of a social statistician. A course in mathematical statistics is an excellent thing for a student, but from the viewpoint

of the sociologist and the social worker it is pedagogically inadequate. Prolonged practice in the application of statistical methods to social data is essential to develop the thought habits necessary in a special field. This book uses illustrative material which centers the attention of the student on a problem of sociology or of social work. Statistical methods are, then, tools with which he may work and are means which may be employed to answer questions about sociology and social work. The introductory course in statistics contemplated in the preparation of this book is one which introduces the student to quantitative aspects of sociology and social work. The author believes that this viewpoint has a distinct advantage in the training of a social statistician over the viewpoint that one kind of illustrative material is as good as another.

This text is intended to provide for a two-hour course throughout the year. The materials for exercises given at the end of each chapter (beginning with Chapter V) are to be used as a basis of practice in using the particular methods under consideration. Teachers of social statistics will usually want to introduce some exercise material which they have found particularly good or which directs the attention of the student to social data in his own city or state. Thus, the book does not prevent considerable latitude in the choice of materials for laboratory practice, while at the same time it places at the disposal of the instructor materials which he may use at his own discretion. If the course happens to be a three-hour course throughout the year, the author has found that the materials of this book can be used satisfactorily by making use of special studies by students. It has been the practice of the author to use as much of the year as is necessary to cover the methods discussed in the book and then to plan special problems for statistical analysis by the students. These problems may involve the use of half a dozen of the common methods of statistics. One of the problems the entire class worked upon was the construction of special and general indexes of public welfare work in Indiana. The construction of these indexes involved definition of terms, the consideration and estimation of population changes, computation of rates and averages, determination of weights, and the computation of trends and cyclical variations. If time permits, several such problems may be studied during the year. The student is required to decide what methods are necessary to answer the questions raised about the problem, and then he is expected to interpret the results of his

work. This experience helps to develop the habits of thought and imagination necessary to a social statistician.

Four chapters have been included in this volume which are not usually found in general texts on statistics. They are: Chapter II, "Sources of Published Statistics"; Chapter IV, "Working Out a Statistical Problem"; Chapter XIV, "Vital Statistics" and Chapter XV, "Rating Scales." Unless the instructor gives a lecture on standard sources of statistical material, the student is likely to have an uncertain idea about where to turn when he wants certain kinds of data. For this reason Chapter II was included as a sort of "bibliography" for the student in statistics. Most texts on statistics discuss the procedure required in working out a statistical problem, but the discussion is generally scattered through the book. There is no objection to this, but it seems to the present author that it is desirable to present this subject in a separate chapter so that the procedure may be shown more systematically. A number of monographs have been written on vital statistics, but general texts usually have given only scant attention to the subject. Yet the social statistician is constantly concerned with births, deaths, morbidity, and population. It seems reasonable, therefore, in a book on social statistics to give a separate chapter to the presentation of a few of the methods of analysis used in the study of vital statistics. Rating scales are fairly new as statistical tools, outside of the scales for the measurement of intelligence, but they give promise of much greater importance to the social statistician in the future. It was felt that the student should become familiar with the nature and possibilities of rating scales in sociology and social work. Hence, Chapter XV was devoted to this subject.

The author of a book on social statistics is inevitably indebted to a great number of his colleagues, known and unknown. Most of all my thanks are due to Professor F. Stuart Chapin, Editor of Harper's *Social Science Series*. Professor Chapin has read all of the manuscript, and in conference and by letter has offered valuable criticisms and constructive suggestions far too numerous to mention in detail. He is entitled to much of the credit for whatever merit the book possesses. I wish to express my appreciation to Professor Robert E. Chaddock, my former teacher in statistics, who stimulated my interest in social statistics and whose clear thinking in his writings has been a constant inspiration to me during the years since I sat in his classes. My thanks are due to Professor Charles R. Metzger, of Indiana University, and to Miss

R. Elizabeth Cox, my secretary, for checking the computations in the book and for assisting with the proofreading. I also want to express appreciation to Professor U. G. Weatherly of Indiana University, for his interest and encouragement during the time the manuscript has been in preparation; he has helped to clarify my conception of the function of statistical methods in sociology by kindly and philosophic criticism.

Acknowledgment is here made to the Johns Hopkins Press for permission to summarize extensively from Schmeckebier's *The Statistical Work of the National Government*; to the University of Chicago Press for permission to quote at length from Thurstone and Chave's *The Measurement of Attitudes*; to Houghton Mifflin Company for permission to reprint the tables of logarithms in Kuhn and Morris' *Mathematics of Finance*; to George Routledge and Sons for permission to use considerable material from Dorothy S. Thomas' *Social Aspects of the Business Cycle*. For aid in the assembling of statistical material for illustrative purposes and for the use of official reports, I wish to express my gratitude to the Indiana Board of State Charities and the Indianapolis Family Welfare Society.

The author wishes to emphasize the fact that he assumes full responsibility for the shortcomings of this volume. Those who have given advice or have assisted in other ways are in no way responsible for its weaknesses.

Indianapolis, Jan. 2, 1933 R. CLYDE WHITE

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

Part One

INTRODUCTION

CHAPTER I

Social Problems and Social Statistics

I. STATISTICS AS DATA AND AS METHOD

SOCIAL statistics *are* data which occur in human society, and social statistics *is* a scientific method. The worker in the social sciences is concerned with both the data and the method. Great masses of social data are received, tabulated, and filed by public and private agencies every year, but up to the present time relatively little systematic use has been made of these collections either for scientific or for administrative purposes. Social statistics as a method of analysis leading to understanding and control is in about the same stage of development as accounting in business was fifty years ago, when the old-fashioned bookkeeper recorded receipts and disbursements, made a balance sheet and called the matter closed. Today accounting requires the recording of facts which its bookkeeping predecessor would have excluded as irrelevant, because accounting is now concerned with unit costs, rates of production, sales per employee, capital depreciation, gross profits, net profits, etc., as interrelated factors which are of primary importance to the success of business. When social statistics analyzes the data recorded by an agency, of whatever sort, from every point of view to determine the effectiveness of the institution in the light of its own reports, it is doing what might be called social accounting. The business accountant records facts, and then applies statistical methods, suited to his purpose, to appraise the business. Too often social agencies, public and private (and here the educational system is regarded as a social agency), record many facts, assemble them in tables, publish or file the assembled data, and carry the work no further. But this is just the point at which the serious work of the social statistician becomes interesting and takes on significance.

Effective statistical work in social institutions requires adequate reporting of essential facts, and then the continuous and systematic analysis of these facts. The United States Bureau of the Census is maintained as a fact-collecting agency, and its primary responsibility ends with the collection, tabulation, and publication of these facts, though actually the Bureau analyzes some of its own material and occasionally issues monographs of first-rate importance. The latter, however, is a secondary function. The Bureau is not a functional agency in the same sense that state departments and city bureaus are. A state department of public welfare, a city park board, a community chest, or a school board exists to carry on definite service to the state or community. Its function is primarily administrative, not fact collecting. But it must have facts upon which to base the policies that underlie efficient administration, and administration will be much less efficient if these facts are not the subject of continuous analysis by a competent statistician. If the annual reports of departments of public welfare and school boards consisted in part of careful analyses of the data reported, they would make exciting reading for the public and would enlighten the administrators on many points. The great masses of official and quasi-public data assembled every year will yield up their meaning only after careful study; they are too complex to be interpreted by rule-of-thumb or impressionistic methods, such as are now commonly in vogue among administrators.

But social statistics requires more than periodic reports and analyses, if it is to perform for social science and social administration a function comparable to experimentation in the natural sciences. Hand in hand with reporting facts goes the judgment as to what is significant. Statistical data are currently or periodically collected to afford a measure of the magnitude of the problems dealt with and to guide the administrator in the direction of greater efficiency and social effectiveness. Other data which bear upon causation may be equally important, if control over conditions is sought. Consequently, it may be asserted that a social statistician must know the field of his operations as well as statistical methods —should even be master of his field of interest before he is concerned with statistics. A mathematician possessed of the most refined statistical technique could not make a significant analysis of crime data unless he had studied crime and learned what factors are probably significant in its causation, control, or prevention. The social statistician must know the subject which is to occupy his in-

terest and must employ his statistical technique to extract meaning from it and to measure trends, variations, and relationships.

It is the purpose of this chapter to indicate specific uses of statistical methods in the field of social problems. The presentation will of necessity be brief, but it will cover in summary form the following points under each problem discussed: (1) data relating to the occurrence of the problem in time, place, and population group; (2) data concerning the magnitude of the problem; (3) data concerning administration and its efficiency; (4) data concerning possible causes; (5) data concerning social control of the conditions.

2. GENERAL EDUCATION

The systematic transmission of the culture of the adult generation to children is the problem of the free public schools. The culture which the schools attempt to pass on includes knowledge, skills, and attitudes. It is the most gigantic social problem with which each generation has to deal, and one which comprehends every child born or brought into the United States; it is the social problem for which the most elaborate machinery has been devised. Tons of paper bearing educational statistics are filed every year. For many purposes these data would be inadequate and would have to be supplemented, but they are adequate for other purposes if they are analyzed and have the juice squeezed out of them. These data bear upon social problems other than education, because the assimilation and utilization of culture may have manifold effects.

Education is the concern of every citizen, urban and rural, rich and poor, educated and ignorant. The problem of insuring every child a minimum of free public education is tremendous. All large cities in this country now conduct their schools for nine or ten months each year, but in rural communities this long period is often not achieved. The property per capita available for taxation is lower in rural communities. There is also a wealth differential among urban communities. Communities with relatively low per capita wealth must levy high taxes, or they must have state or federal aid. Almost all states have a public school fund in which local communities share according to the number of children of school age, but this kind of aid does not equalize the opportunity of all children to receive education because many communities are handicapped by low per capita wealth. Only a system of special

state or federal aid can equalize educational opportunity. Some states provide such special aid. In the determination of the amount of aid required and the places requiring it statistical information is indispensable, and competent analysis of such information is no less necessary. It is the problem of state departments of education to determine where special aid is required to bring local schools up to standard achievement. The communities suffering from inadequate schools change from one year to another. Children have varying ability to profit by school education; consequently, within a city or a rural county there arises the problem of provision for typical children. Surveys have shown that whole counties may have a disproportionately large number of mentally handicapped children. Hence, the composition of the population is an important factor in school administration. Only continuous study of such economic and social problems by one trained in the methods of research can insure its efficiency.

The magnitude of the national education problem is indicated by the fact that in 1930 there were approximately thirty-five million children between five and twenty years of age. To educate this number requires about a million teachers, besides many thousands of administrators, research people, and clerical workers. In the school year, 1927-28, there were in continental United States 257,251 public school buildings, and the value of school property was $5,423,280,092.[1] Such a stupendous undertaking can go forward with any degree of satisfaction to the public only on a basis of sound statistical data and careful study and planning. Of course the administration of this vast institution is divided among states, counties, and local communities, but even they must rely upon statistics for guidance. Where state financing and supervision are factors, the quantity of data required is large, and the difficulties of interpreting them are great.

Administration of a school system, whether by state, county, or city, requires an understanding of statistical methods and the ability to draw conclusions and formulate policies from masses of data. Periodically a city school administration has a survey made to take stock of its routine and social efficiency. Such sporadic surveys are implied confessions that the school system has not collected currently all the data necessary for statesmanlike administration, or that competent statistical service was lacking, or both. City school systems are making increasing use of statisticians, but some of the

[1] World Almanac, 1931, p. 402.

annual reports still bear witness to the lack of appreciation of the value of such service in administration. This, however, is entirely aside from the broader social questions. The relation of schools to delinquency, to utilization of leisure time, to health education, to morbidity, etc., is a fact in which the community is vitally interested, but, when such problems are considered, they are usually analyzed in some survey report instead of being analyzed from routine research, which is of more value to the community. Even a city school system which has a good social service department makes no annual analysis of its data; the records are filed, and only individual cases ever come to light to affect administration—in spite of the fact that the first principle of good statistical procedure is that a generalization can be made only from a consideration of all cases or of a representative sample.

Public education is a social problem because group conflict and differences in economic status and individual differences in ability exist in every community. Within the school system administrative problems may appear to be purely professional matters, but the way in which they are worked out affects the education of the child and, consequently, the community. The good school administrator is interested in every social relation of the school and he can judge the trend of development in these relations only through statistics currently analyzed and interpreted. He can gain control over developmental tendencies either by shrewd guessing or by scientific study of the social and technical facts. Shrewd guessing is still the more common practice, but in some school systems steps are being taken to direct public education on the basis of continuous statistical analysis of facts.

3. EMPLOYMENT

Employment which is economically useful for all able-bodied adult members of the population is an American ideal which goes back to the earliest colonies. It becomes a social problem because American democracy desires full opportunity for each individual to earn his own and his family's way in the world in the occupation for which he is best fitted, and also because our economic system is such that many men and women have to endure involuntary unemployment at various times. This may be seasonal in a certain locality because of the nature of the occupations available, and this kind of unemployment tends to recur every year at the same time. A business depression causes what is known as cyclical unemploy-

ment, when workers are laid off for months at a time. Cyclical unemployment is usually national or international in extent. Rapid changes in machine production and in administrative organization create what is called technological unemployment. This form may occur any time in a factory, a department store, or on a farm, if labor-saving machinery or new administrative devices are introduced. To keep people employed at productive labor is the positive way of stating the problem of unemployment; relief for the unemployed and the prevention of unemployment is the negative way, but the latter is the more common approach to the problem. Many social problems for the individual, the family, and the community arise in the wake of unemployment.

Although seasonal and technological unemployment occurs every year and cyclical unemployment every six to ten years, statistics are not available to indicate the magnitude of the problem. The community has given little attention to the first two types, except to maintain agencies for charitable relief; but when a serious business depression occurs, the distress caused by unemployment focuses a great deal of attention on the problem. Yet in previous depressions the estimates of the number unemployed in the country have varied by millions, a fact which merely emphasizes the dearth of statistics bearing upon a problem that can be dealt with adequately only when reasonably dependable data are available. The United States Bureau of the Census undertook a census of the unemployed in 1930, but the returns were so much in dispute that it is not known whether this census actually indicated the magnitude of cyclical unemployment in April, 1930. A few cities have made estimates which may be fairly accurate. The Indianapolis Commission for Stabilization of Employment estimated that the percentage of the employable population in that city who were employed declined from 97.2 per cent, March 31, 1930, to 78.1 per cent, December 31, 1930. Some of those who were unemployed December 31 were undoubtedly out because of the normal seasonal drop in employment, but the data available are insufficient to compute a seasonal index of employment which should be deducted from the total unemployed, to arrive at the number out of work because of the depression. However, cyclical unemployment usually involves millions of families, and the magnitude of the problem is reflected in the rising amounts paid out for charitable relief. But to the social statistician the important point is that statistics are inadequate in quantity and dependability.

Employment administration as a public responsibility is only beginning to receive attention. Some of the large cities operate free employment exchanges; these are of great value in diminishing the period of idleness of the individual worker who is unemployed on account of seasonal and technological changes, but of small utility in a general depression. A national system of free employment exchanges is probably coming into existence, but the effectiveness of its administration will depend as much upon current statistics of employment in the locality, the state, and the nation as upon organization and trained personnel. Efficient employment administration requires detailed statistics, currently collected, concerning seasonal variations and technical changes in all important businesses in the city. Cyclical unemployment can be dealt with effectively only when the organized community has sufficient current information to anticipate increasing general unemployment some time in advance, and can promptly set in motion public works, emergency work, and other relief measures of such comprehensiveness that the volume of unemployment will not demoralize the community and the families of the unemployed.

While the causes of seasonal and technological unemployment are fairly well known, there is much debate concerning the causes of cyclical unemployment. Seasonal lay-offs occur because the buying habits of the public concentrate purchases of certain commodities at particular seasons, because raw materials are available only at certain times and may be perishable, because outdoor work in the winter is difficult and inefficient, because second-line industries sell their products to primary industries which have seasonal variations, and because some producers are in the habit of speeding up for a part of the year and slowing down at other times. An understanding of these conditions and methods of removing them requires more information than is now available, and more intensive and comprehensive analysis. A single new industry of large proportions, such as the automobile industry, so affects the whole economic system that it is necessary for every community to have current statistics bearing on the problems of unemployment in order to know the causes of a given condition.

Social control of the conditions leading to unemployment or to the alleviation of its effects is not far advanced. Stabilization of employment through the efforts of corporation executives has reduced the number of seasonal lay-offs in particular businesses and offers one way for further advance. The free employment ex-

change is about the only agency yet developed which can readjust workers in new jobs when they are thrown out of work by technological conditions. No control over conditions leading to cyclical unemployment exists and here only relief measures are available. England and Germany have set up systems of unemployment insurance, benefits from which are available to anyone who is unemployed and cannot find work provided he is in the categories of the workers insured. If such a measure should be adopted in the United States, it would at the very outset require vast information to work out th. plan on a sound actuarial basis. Once put into operation, it would take a small army of statisticians to keep up with the collection and analysis of data. Unemployment, more than any other social problem, requires the use of statistics and statistical methods, if it is to be handled in a statesmanlike manner.

4. POVERTY

"The condition of poverty obviously attends every person who habitually lacks the means to sustain himself on such a footing of physical fitness as will enable him to carry on effectively for himself and his legal dependents. Such a person may not be in abject want and yet be in poverty. He may be a laborer whose weekly wage is barely sufficient to sustain life, leaving no margin for advancement. He is not in danger of immediate death from starvation, but he lacks enough to maintain a permanent and reasonable standard of physical fitness."[2] Poverty is thus defined in terms of physical health. Of course, economic standards of living change, and probably the concepts of physical fitness vary also. But in defining poverty in terms of physical health a more objective approach to the problem is insured. Poverty is the usual condition of what is sometimes called the "submerged tenth." According to this opinion, if the economic status of any large number of people in a given geographic area were known, it would include those in poverty as about ten per cent of the total. Where the foreign-born and the Negroes constitute a large proportion of the population, the incidence of poverty is probably much greater. Its occurrence is more obvious in certain districts of large cities than in rural areas, but this may be only apparent because so many poor people live close together in cities, the laboring population tending to live as near their work as possible to save car fare and because house rents

[2] Kelso, Robert W., *Poverty*, p. 3. New York: Longmans, Green & Company, 1929.

are low in areas contiguous to industry. Poverty is more common among industrial laborers and farmers than in any other occupational group.

That the problem of poverty is one of great magnitude is indicated by the large sums of money spent every year in poor relief, and the large numbers of persons receiving such relief. In February, 1929, fifty private relief agencies distributed $514,007 to 21,069 cases—in the majority of these a "case" is a family; and the same agencies distributed $3,986,958 to 221,550 cases in February, 1932.[3]

Current indexes of general business conditions in February, 1929, were a little above normal for that month, and considerably below normal in February, 1932. Possibly the difference between the number of relief cases handled by these agencies is some indication of the numbers of people who live in poverty but who do not require charitable aid except in times of business depression. Another indication of the number of the poverty-stricken requiring aid is given by reports of public poor relief in certain states. Between April 1, 1928, and March 31, 1929, Massachusetts spent $12,851,771.51 for the aid of 149,523 persons. For each thousand of the population 37.21 persons received public aid, or 3.72 per cent.[4] In Indiana, for each thousand of the population (estimate), 43.05 persons, or 4.31 per cent, received outdoor relief in the fiscal year ending September 30, 1929, and in poor asylums 2.07 persons (includes holdovers, new admissions, and readmissions) per thousand population were given aid during the fiscal year ending August 31, 1929. The total cost of these two kinds of poor relief was approximately $2,862,500.[5] Besides these public charities illustrated by Massachusetts and Indiana, there are many private agencies giving relief, and many other agencies give services to people unable to pay for them. The figures mentioned here suggest the magnitude of the problem of poverty, but they do not give any exact measurement. A problem so great and so expensive should warrant more complete statistical records and a more systematic analysis of the data.

Poor relief is administered as a dole system by public agencies;

[3] Published reports of the Russell Sage Foundation.

[4] *Annual Report of the Massachusetts Department of Public Welfare, 1929*, pp. 132-134.

[5] *Indiana Bulletin of Charities and Correction*, No. 182, pp. 203, 204, 302, 303, and Nos. 183-184, p. 365.

the exceptions in which the principles of social case work are employed are too few to make much difference. The private relief agencies are increasingly giving relief only as a part of the process of rehabilitation, and it is these agencies which have seen the importance of fuller records and of employing statisticians in their work. Public relief of poverty, as it is now administered, is generally believed to contribute to pauperism. Whether it does or does not is a matter to be determined by more complete data and their analysis. In Indiana, the trend of public poor relief in proportion to population, for the past thirty years, has been upward. Does this reflect a more liberal policy on the part of overseers of the poor? Or does it reflect a growing class of poverty-stricken citizens? Statistical research would help to answer these questions.

Information on the causes of poverty exists in the records of public and private relief agencies, but it has not been studied scientifically to any important degree, although case studies and some efforts at statistical summary and analysis have been made by individual social agencies. But the question narrows down to a judgment as to whether poverty is due primarily to personal inadequacy in modern civilization or to defects in economic organization. Low wages in large families is undoubtedly a factor, because a business depression sends to relief agencies many persons who do not require such aid under ordinary circumstances. Low mentality seems to play an important part as a cause of poverty, and personality disorders come in for consideration. Disasters, accidents, and illness precipitate people into poverty. No doubt, the relative importance of these factors varies in different localities. Because this is true, continuous local records and their systematic analysis are fundamental to a comprehensive understanding of the specific causes of poverty.

Control over the conditions which lead to poverty waits upon more certainty concerning these conditions. Increased wages and stabilization of employment might improve the situation; the early detection of physical defects and peculiarities of personality might help in the control of two groups of cases; segregation or sterilization of the feeble-minded would prevent this class from rearing families in poverty. But all such efforts at control depend upon more careful scientific work than has yet been done in the field of social problems and social work. Relief, unemployment insurance, pensions, and made-work are palliatives to minimize the distress

of the victims of poverty; they are not means of control over causes.

5. OLD AGE

Dependent old age is increasingly a social problem. The proportion of aged persons varies in time and place and according to ethnic composition of the population. In 1850 the percentage of males sixty years of age or over was 4.0 and of females 4.2, but in 1920 the percentage of males in this age group was 7.4 and of females 7.5. The relative number of the aged in the United States has almost doubled in seventy years. The percentage of persons sixty-five years of age or over in 1920 varied from 3.4 per cent in the West South Central States to 5.8 per cent in the New England States. By ethnic composition the percentage of persons sixty years of age or over varied as follows in 1920: native white parents, 8.1 per cent; foreign-born parents, 4.9 per cent; mixed parents, 6.1 per cent; foreign-born, 14.9 per cent; Negro, 5.1 per cent. Thus it will be seen that the determination of the occurrence of aged persons is a statistical problem itself, and, when it is related to social factors, dependent old age becomes an exceedingly intricate one.

To a considerable extent dependent old age as a problem to the nation and to local communities arises from income inadequate to permit saving for old age, to financial inability of adult children to take aged parents into their homes, to the tendency of employers to discriminate against older men, and to the fact that the percentage of the population above sixty years of age is steadily increasing. The magnitude of the problem is suggested by the percentages just given. It is further emphasized by the actual number of persons sixty-five years of age or over in the United States in 1930, which was 6,633,805. The problem is not materially reduced when it is recognized that about half this number are women, since they must be taken care of either in their own homes or somewhere else, and it is a fact that their husbands are finding it increasingly difficult to find employment, if they are wage earners.

That part of public administration which deals with the aged is concerned almost exclusively with relief. In sixteen states old age pension systems have been introduced. In some states the minimum age for eligibility is sixty-five and in others seventy, all states providing that persons eligible by age are not rendered ineligible

by the possession of more than the maximum of property allowed. That old age pension systems are expensive is indicated by the fact that New York State spent about $12,000,000 for 1931, the first year of its operation. Other relief of the aged poor is left to the charitable agencies and to the poor asylums. In the private relief agencies rehabilitation of the aged is undertaken, and case-work treatment seems to give promise of good results. Public outdoor relief agencies simply dole out relief without any effort at constructive work. The poor asylums are in fact custodial institutions in so far as the permanently incapacitated old person is concerned, and the private homes for the aged are of the same general character, though they are generally better managed and more comfortable. The poor asylums, of course, do not restrict admission to elderly persons, but in Indiana in recent years over two-thirds of the poor asylum population have been sixty years of age or over. Little effort is made any where in the country outside of the private case-working agencies to prevent dependence in old age. Prevention is left to the individual or to his family. Occupational adjustment might be possible on a much larger scale for the able-bodied person past sixty. Much more careful study of old age relief and of preventive possibilities needs to be undertaken both by departments of public welfare and by private agencies.

The causes of old age dependence are believed to be many. They include illness, mental disorder, mental deficiency, personal improvidence, insufficient income in productive years, criminal behavior in earlier life, and the disinclination of employers to take on elderly people. The relative importance of these factors in different localities is not known, and control of the conditions which lead to old age dependence requiring charitable relief or a pension cannot go far until the problem is better understood. The states which have adopted old age pension systems should become laboratories for the study of old age problems. It may be found that pensions encourage dependence and remove important incentives to self-maintenance or to coöperation in a plan of prevention. Careful records and thorough statistical analysis are indispensable prerequisites to the solution of this problem.

6. DEPENDENT AND NEGLECTED CHILDREN

A dependent child is one whose parents are dead or are incapable of taking care of him and whose near relatives cannot assume responsibility for him. A neglected child is one whose

parents or near relatives do not give him the care he needs but who may be financially able to do so. In such cases either a private children's agency or the state undertakes the care of the child. No social problem receives more attention than that connected with children. This is true because the thought of children inadequately cared for arouses sympathy and immediate action, but also because as a practical matter the children who are neglected or lack the elemental necessities of childhood soon perish or grow into adulthood with numerous handicaps. If a child cannot be reared satisfactorily in his own home, society reserves the right to make him a ward of the state and to supply as much as possible of what is lacking.

In 1920 nearly one-third of the population of the United States was under fifteen years of age. This group furnishes the problems with which child welfare efforts are concerned. Dependency and neglect are two of the most common problems. Probably they occur more often among the foreign-born and the Negro population than among other groups, though reliable statistics for the country as a whole are not available to show the exact condition. They seem to occur more often in the city than in rural communities, but this may be more apparent than real since facilities for detecting these conditions are more numerous and better organized in cities. Some geographic divisions of the country report much larger rates of dependency and neglect than others, but in the absence of statistics to prove the point this fact may be assumed to reflect differences in standards of child care rather than differences in the rate of occurrence of the problem.

As compared with other social problems, the magnitude of the problem of dependent and neglected children can only be suggested. A report of the Bureau of the Census indicates something of the situation.[6] For every 100,000 of the total population of the United States in 1923, 198.7 dependent and neglected children were reported; but in New England the rate was 353.0, and in the West South Central States it was only 98.7. On February 1, 1923, there were 148,979 children in institutions for dependent and neglected children or under the supervision of these institutions. On the same date 339 child-placing agencies reported 52,979 children under their care. Children are continually coming under the care of such agencies, though, of course, other children are

[6] *Children Under Institutional Care, 1923.* Bulletin of the United States Bureau of the Census.

continually being released. Between February 1 and April 30, 1923, 1,558 institutions reported that they had received 9,198 children, and 339 child-placing agencies received 7,181 in the same period. It is estimated by the Bureau of the Census that another group of children numbering about 121,000 is under care of mothers' pension administration. Besides these agencies, there are day nurseries and certain institutions which receive pre-delinquent and mildly delinquent children on the same basis as dependent and neglected children.

These figures suggest the magnitude of the problem, but they do not indicate the degree of efficiency in administration attained by institutions and child-placing agencies. Such data as a whole are entirely lacking. A few studies of limited extent have been made,[7] but they do not reflect the results obtained throughout the country. That is a technical problem requiring more data than are now available for analysis. Much research has been, and is being, done to determine the best ways of handling dependent and neglected children, but in the main this is case study which throws light only upon methods of individual treatment. The larger problem of the interrelationships of dependency and neglect with other social factors has received much less attention, chiefly because technically it is a statistical problem.

The proximate causes of dependency and neglect in individual cases are usually known, but why the social order should produce such pathological conditions is still far from a scientific answer. This question becomes more difficult, when it is known that the number of such children under care in proportion to population seems to be increasing. The determination of whether this indicates growing pathological conditions or more active response to the needs of children is a first-rate problem for social research. No control in the sense of preventing the conditions which give rise to dependency and neglect is possible until much more research has been done. Statistical records and statisticians are essential to the solution of much of this problem. Years will be required for the accumulation of facts. Current statistics gathered by departments of public welfare should be studied as they come in, but so many factors are involved that a full understanding will be achieved only after observation of many annual series over a period of years. Time is itself an important factor, and that is

[7] See Van Theis, Sophie, *How Foster Children Turn Out*. State Charities Aid Association of New York, New York, 1924.

perhaps one reason why the best results in the study of the problem will be obtained by research workers who are regular members of departments of public welfare and who serve with an indefinite tenure of office.

7. DIVORCE

Divorce is becoming easier, and consequently more important as a social problem. The more applications for divorce there are, the more time the courts have to give to this type of litigation. When a married couple seeks to dissolve their marriage relations, the public is concerned not merely with the fact that a decree of divorce may be issued to the husband or the wife. Other matters of public importance are involved: the disposition of children, the division of family property, and the socio-psychological effects on the parties concerned. More than a third of the divorce cases involve children. It is generally believed, though not proved conclusively, that children are socially handicapped if their parents are divorced; and there is considerable evidence that behavior problems develop in such children more readily than in children living with both parents. Slightly more than one-third of divorces are granted in the first five years of marriage, when there are no children or the children are small. The occurrence of divorce varies among states according to the liberality of the laws. No divorces are granted in South Carolina, but in Nevada they are granted freely. Texas, with a population only half as large as New York, has three times as many divorces.[8] Divorce occurs much less frequently among Catholics or Jews than among Protestants. Economic conditions so affect the divorce rate that in times of depression there is a distinct drop, while in prosperous years the rate shows a marked rise.

The magnitude of the social problem created by divorce can only be suggested. In 1928 the Bureau of the Census reported that 195,939 divorces were granted, and the 1920 census showed that there were 508,588 divorced persons who had not married again. It is probable that the majority of persons who get divorces remarry. So the census figures do not indicate the number of persons in the population who have at some time been divorced. The number is much larger than that for divorced persons remaining un-

[8] See *Marriage and Divorce, 1928*. Bulletin of the United States Bureau of of the Census.

married. Ogburn estimates that 0.7 per cent of the total population was divorced in 1920.[9]

Judicial divorce statistics are inadequate. The courts are concerned with individual cases as they come through; certain information is obtained and filed. Statistical reports cover only a few items which are quite insufficient for either administrative or social purposes. The courts have not seen the value of statistical studies of their work, which would enable them to compare accurately the grist of one year with that of another. A research project in Ohio and Maryland is now being carried on by the Institute of Law of Johns Hopkins University for the purpose of determining what statistics are most valuable for judicial administration and for social reporting. When this question is answered, it still remains to get the plan of reporting adopted by courts and, most important, provision made by the judicial system for competent, current analysis of the data.

The causes of divorce are not those alleged in the legal grounds for divorce. The complaint is made in a form which the complainant believes will meet the requirements of the law, but the specific circumstances which led the parties to decide to dissolve their marriage relations do not often appear. In a large proportion of the cases the causes lie in the peculiar personalities of the marriage partners. These factors are highly indefinite and difficult to express in statistical units. If the socio-psychological factors can ever be represented adequately by more objective factors, it may be possible to make a comprehensive statistical study of the causes of divorce. At present such a study cannot advance far. Social control of divorce depends upon the law and the attitude of leniency or strictness shown by the judge. More adequate statistics of divorce, which would show clearly its effects, would be some guide to the future modification of the law. These have yet to be developed.

8. CRIME AND DELINQUENCY

Briefly, crime is a violation of the law. Delinquency is a term usually applied to young offenders, and may be a violation of the law or may be anything that might lead to an overt violation. Next to the problem of unemployment, crime is probably the most ex-

[9] Groves, E. R., and Ogburn, W. F., *American Marriage and Family Relationships*, p. 360. New York: Henry Holt & Company, 1928.

pensive social problem confronting the nation. Efforts have been made to locate the occurrence of crime specifically in time, place, social strata, age and sex groups, and ethnic groups. Some success has attended research in certain cities in defining the geographic centers from which the bulk of crime springs; in all cities so far studied the concentration is just outside the main business district and in certain outlying industrial districts. The seasonal variations are less well defined, though certain types of crime appear to have regular ups and downs. The age and sex distribution is rather well known. The ratio of males to females on January 1, 1930, was about twenty-four to one.[10] Among juvenile delinquents the ratio of males to females is about four to one. About one-half of the males and about two-thirds of the females are fifteen to seventeen years of age.[11] Statistics on the distribution of crime by social strata are insufficient to draw a conclusion, but it is probable that a disproportionately large number of criminals and delinquents come from the lower economic classes—those who would be unskilled or semi-skilled workers. The foreign-born and the native white of native parentage, when age and sex are held constant, probably show the lowest rates, while the Negro and the native white of foreign parents show higher rates. A few particular foreign-born groups appear to have high rates, however. Altogether, there is still a good deal of statistical work to be done in locating the occurrence of crime and delinquency.

The magnitude of the crime problem, as reflected by estimates, is staggering. Much work has been done to produce standard criminal statistics, but much remains to be done. The number of convictions by courts furnishes the most reliable statistics, but convictions are so small a percentage of crimes that they do not accurately represent the volume of crime. On January 1, 1930, there were 120,496 inmates of penal and reformatory institutions, and the commitments in 1930 were 78,866.[12] Furthermore, a great many prisoners were on parole from the institutions, and many more convicted criminals were on probation. Hence, institutional statistics are inadequate as a measure of the volume of crime. Similar statistics are available for juvenile delinquents. On January 1,

[10] *Prisoners*, 1930. Report of the United States Bureau of the Census.

[11] *Children Under Institutional Care*, 1923. Report of the United States Bureau of the Census.

[12] *Op. cit.*

1923, 25,233 juvenile offenders were in institutions, and about that many more were admitted during the year.[13] These figures, even more than those for adults, fall short of reflecting the real situation, because many more juveniles than adults are put on probation. Aside from the social losses to the country through this volume of crime and delinquency, the cost of maintenance of police systems, courts, and institutions is stupendous, and to these costs must be added the destruction and theft of property by criminals.

The administration of the courts, the police systems, and the institutions is the kind of social problem that calls for ample carefully made statistical records and a systematic analysis. During the last few years the federal government and many of the states have appointed crime commissions to survey the situation. Such sporadic surveys have been made before, but they accomplish little other than tightening of the law and for a short time directing the attention of the public to crime. Administration requires continuous factual records, just as a corporation requires continuous accounting, and efficient administration requires at best annual, systematic analysis of the work of the year—not a popular report for the press but a report that is technically as competent as that presented by the officers of a corporation to the board of directors. Probably no court, police system, or institution in the country has an equally competent annual analysis of its work. Changes in the law, tightening court procedure, and more vigorous police efforts are usually the extent of the effects of a crime survey. Administrative efficiency requires police, judicial, and institutional accounting of a high order.

Study of the causes of crime has hardly arrived at the stage of statistics except by indirection. Criminals seem to come from an economic class whose income is for the most part in the lowest quarter; they live prevailingly in neighborhoods which are undesirable to most people for residential purposes; there seems to be a disproportionately large number with low-grade mentality and with mild or serious mental disorders; thwarting of personality in childhood seems to be a causal factor; racial and ethnic discrimination seems to appear as a cause in some crime and delinquency. Studies of prisoners by Glueck, Vold, and Burgess have led to the development of scales, constructed from social background data, which indicate the expectancy of success or failure

[13] *Op. cit.*

on parole.[14] If after sufficient use such expectancy tables prove to be a reliable guide, then it would seem that the chief causes of antisocial behavior have been found. Control over the conditions which develop criminals and over the reconstruction of behavior depends upon further statistical study of this kind. Administrative efficiency and control will very likely advance together.

9. BIRTH AND DEATH RATES

Birth and death rates are biological facts, but they are of importance to the scientific study of almost every social problem. From the rate of increase in population due to the difference between the number of births and the number of deaths and between immigration and emigration school administrators can estimate the amount of equipment, the number of buildings, and the teaching staff that will probably be needed several years in the future. Specific birth rates vary in different social and ethnic groups, and they vary according to the age and sex composition of the population. Specific death rates in age and ethnic groups vary widely and are important in the study of public health work, employment, and poverty. Death rates vary in certain geographical areas, even though the rates may be computed for a standard population. Births and deaths are universal phenomena, and wherever social problems exist they are factors that need to be taken into consideration.

Births and deaths have been recorded for many years, but even now there are areas of the United States not included in the "registration area." Many counties do not have health officers, and in these counties reports of vital statistics are incomplete or wholly lacking. But the problem of getting data for computing birth and death rates is not as great as the problem of obtaining some other kinds of data. Reporting is fairly well standardized for births and deaths, and crude birth rates and general death rates are reasonably reliable. This is not true of specific birth and death rates, however, because their computation requires detailed information regarding age, sex, and ethnic composition of the population which are available for the country only every tenth year, when

[14] Glueck, Sheldon and Eleanor T., *500 Criminal Careers*, New York: Knopf, Chap. 18, 1930.

Vold, G. B., "Factors Entering into the Success or Failure of Minnesota Men on Parole," *American Sociological Society Papers*, May, 1930, pp. 167-169.

Bruce, Harno, Burgess, and Landesco, *Parole and the Indeterminate Sentence.* Department of Public Welfare of Illinois, 1929.

the census is taken—a few states take a census in the middle of each decade. Consequently, the composition of the population in intercensal years has to be estimated, and birth and death rates computed on the basis of these estimates are subject to considerable error.

The administrative efficiency in the collection of vital statistics depends largely upon the public health organization of the several states. If state boards of health are seriously interested in vital statistics, they can gradually build up a satisfactory system of reporting. Because the use of vital statistics is of long standing, boards of health are more likely than other public departments to employ their information in the study of causes and of control. They usually employ a statistician whose main business it is to collect and analyze vital statistics.

10. MORBIDITY

Morbidity is a major social problem and is the cause of many other social problems, particularly those involving inadequacy of income. Illness occurs to every human being at some time in his life. Preventive medicine is aimed at reducing the frequency of disease and in some cases at its virtual elimination. Less is known about the occurrence of morbidity than of mortality. Even communicable diseases are not reported to a central authority in all parts of the country, and acute diseases of a noncommunicable type are never reported unless they are treated in public hospitals. Private hospitals and physicians keep records for their own purposes, but these are not assembled in a central collecting agency so that they may be studied.

In point of magnitude morbidity is one of the most important social problems. The economic loss due to loss of time from work and to actual medical costs is stupendous. Dr. Louis I. Dublin estimates that there are 150,000 physicians, 50,000 dentists, 150,000 nurses, and 100,000 other employees concerned with the care of the sick.[15] The income of these groups and the costs of hospital service and medicine amount to about two billion dollars a year, or about 3.5 per cent of the national income. If the loss of time from work were added to the direct costs of illness, the total bill for illness would be much larger.

Statistical study of noncommunicable diseases has been meager

[15] Dublin, Louis I., *Health and Wealth*, Chap. II. New York: Harpers, 1928. Other statistics in this paragraph are taken from the same reference.

up to the present time. Professor George C. Whipple said in 1923: "It is much to be regretted that at the present time there is no adequate way of getting the facts in regard to sickness in the community due to diseases which are non-reportable. Sickness surveys are sometimes made, but they give only the facts at a given date, and are, moreover, very expensive to make. Hospital records help a little, the examinations made by the life insurance companies help a little, the recent examinations of men for the army have helped a good deal, but some day a more universal method must be devised."[16] The situation has not changed much since Whipple made that statement. Some state boards of health make a commendable effort to collect morbidity statistics, but the results are too inadequate to be of much use for either scientific or administrative purposes. There are no laws compelling physicians to report all diseases; frequently there is no official agency to which they could report. The public has not attached the same importance to reporting morbidity that it has to mortality.

The study of the etiology and treatment of disease is the function of the science of medicine. But the occurrence of disease is a social problem and may properly be the object of study by the social statistician. Even the medical man has made little effort to employ statistical methods as an aid to an understanding of disease; he has been concerned with cases and has not made much use of quantitative studies. The social statistician is concerned largely with environmental data. It is properly his interest to seek more adequate data bearing on his problem and to present the results of his study as a contribution to the knowledge of the causes and control of disease in so far as environmental conditions play a part. Public health officers are obviously concerned with environmental factors. Dr. Thurman B. Rice, of the Indiana University School of Medicine, has found that there is a geographic concentration of goiter in Indiana. This area has been so affected by geologic changes that the iodine in the soil has been leached out. Water in that area lacks iodine content, and food products grown there are deficient in iodine. Goiter is much less prevalent in adjoining counties which have been affected differently by geologic changes. The concentration of a type of disease in any region or population group suggests the presence of environmental fac-

[16] Whipple, George C., *Vital Statistics*, pp. 122, 123. New York: John Wiley & Sons, 1923.

tors. But research of this kind cannot be done extensively until morbidity is more completely reported.

II. INSANITY

Insanity is both a medical and a social problem. As the former, it is receiving a great deal of attention from the medical profession. The case for its social study is equally strong because, aside from the possibility of a social etiology, mental disorder is a complicating factor in many other social problems. Mental disorders are roughly classified as functional and non-functional. The difficulty of drawing a sharp distinction between these two classifications has so far made impossible a judgment as to the relative importance of physical and social causes. At the present time functional disorders constitute much the largest proportion of all mental disorders. "If we accept the opinion that certain neuroses and psychoses are functional," says Professor Ogburn, "and that they indicate a lack of psychological adjustment of man to civilization, then the very great probability of developing in the course of a lifetime a functional psychosis or neurosis certainly indicates a very serious psychological maladjustment between man and his civilization."[17] The problem raised by Ogburn is an important one in social statistics, because the symptoms manifested by an insane person are often so complex that a determination of the definite cause is next to impossible, whereas a statistical analysis of a great many factors in the experience of a large number of persons with mental disorders might result in the discovery of significant correlations. The rate of insanity for different age groups increases with age for both males and females. The occurrence of insanity in different social groups has for the most part yet to be determined.

The magnitude of the problem of insanity, especially its social aspects, can be estimated but is not definitely known.[18] In 1923 there were in mental hospitals 240 patients per 100,000 population over fifteen years of age in the United States. Since many patients recover and are discharged, the number of patients in hospitals represents a disproportionately large number of chronic cases, and does not afford a basis for estimating the incidence of mental disorders in the population. Probably a study of new ad-

[17] Ogburn, Wm. F., "The Frequency and Probability of Insanity," *American Journal of Sociology*, Vol. XXXIV, No. 5, p. 831.

[18] The estimates in this paragraph are taken from Ogburn, *op. cit.*

missions would be more satisfactory as reflecting increase or decrease. New admissions in hospitals in the United States in 1910 were 66 per 100,000 population over fifteen years of age, and in 1927 the rate was 109. This is a marked increase due either to an actual increase of insanity or to more adequate hospital facilities. Ogburn estimates that one in twenty-two boys over fifteen years of age in New York State will probably be a patient in a mental hospital during his lifetime. Using some data obtained in the army medical examinations as a basis for estimating the number of persons in the population who may be afflicted with a mental disorder, many of whom will not be hospital patients, he concludes that in Massachusetts and New York the chances are that one in ten of the population above fifteen years of age will be so afflicted.

Another way of suggesting the size of the social problem of insanity is represented by the financial cost involved. In 1923 there were 153 state hospitals caring for insane patients, with a capital investment of $246,348,925.52—these figures omit twelve other state hospitals. Maintenance in 1927 for the same hospitals amounted to $77,731,015. Thus it will be seen that the costs of insanity are great and constitute a large item in public budgets.[19]

In every state there is some organization for the collection of statistics of insanity, but the facts reported are adequate only for forming an estimate of the volume and cost of insanity and the types of cases in institutions. For statistical work, which would be useful in administration, many more data are required. Usually a hospital draws its patients from certain counties which constitute its district. Frequently a superintendent is impressed by the concentration of cases in a county, for which there is no obvious explanation. More complete statistics of cases correlated with population data might lead to an understanding of the concentration. But to be useful to state hospitals this kind of work should be done currently in each state.

The study of causes of insanity has been confined largely to cases; little systematic effort has been made to apply statistics in a thoroughgoing way. Yet this is a method offering much promise of fruitful work, and it is indispensable for control. The maladjustments between man and his civilization which Ogburn has suggested as causes of insanity must be studied statistically and especially by the correlation technique. In this way it may in time

[19] *Patients in Hospitals for Mental Disease*, 1923 and 1927. Report of the United States Bureau of the Census.

be possible to estimate the relative importance of hereditary, physical environmental, and social environmental factors. The social factors are perhaps more amenable to control than either physical or hereditary conditions. It is, therefore, all the more important that the study of social factors be undertaken.

12. MENTAL DEFICIENCY

Mental deficiency is a broader term than feeble-mindedness. It includes the feeble-minded, but also many others who are less retarded. Dr. Stanley P. Davies quotes the following definition of feeble-mindedness from the Report of the Mental Deficiency Committee of England, 1929: "The only really satisfactory criterion of mental deficiency is the social one, and if a person is suffering from a degree of incomplete mental development which renders him incapable of independent social adaptation and which necessitates external care, supervision and control, then such a person is a mental defective."[20] Others are relatively deficient, even though they may not be classified as feeble-minded. It was once believed that all feeble-mindedness was hereditary; that is, that it occurred only in families where one or both parents or recent ancestors were feeble-minded, but now it is believed that about half the feeble-minded children are so limited mentally because of environmental causes. Mental deficiency is usually indicated by inability to make normal social adjustments because of intellectual limitations. Such persons cannot profit normally from ordinary school training and cannot make satisfactory occupational adjustments except, in some cases of the higher-grade mentally deficient, in unskilled work. Imbeciles and idiots, the two lowest grades of mental defectives, require constant care and often cannot attend to their simplest personal needs.

The number of mental defectives in the population has been variously estimated. Dr. Davies has estimated that there are probably about eight feeble-minded persons per 1,000 population in the United States, which would make about 1,000,000 at the present time.[21] Perhaps twice as many more are deficient in a less degree. The latter constitute a greater social problem than the feeble-minded. They appear with disproportionate frequency among the applicants for charitable relief, the misfits in school,

[20] Davies, Stanley P., *Social Control of the Mentally Deficient,* p. 6. New York: Crowell, 1930.

[21] *Ibid.*

the unemployed, the dependent children, the delinquent, and the criminal. They have greater difficulty than the individual of average intelligence in making all social adjustments.

Three general methods are available for dealing administratively with the mentally deficient: segregation in institutions, sterilization, and care in family homes. Segregation prevents reproduction, if it is permanent, and provides care for the low-grade mental defectives, but it is very expensive. Sterilization definitely prevents reproduction, but it does not solve the problem of care. This method is perhaps better suited to the higher-grade defectives who are able to earn their living. Care in family homes is cheaper than institutionalization and under proper supervision it offers protection to society. Statistical records of administration are not sufficiently complete, and what records there are have been studied much less than they might have been. That is, the bookkeeping for mental defectives does no credit to the administrators.

The causes of mental deficiency are known to some extent. A good many mental defectives inherited their deficiencies and probably carry defective germplasm themselves. Birth injuries, thyroid deficiency in the mother, certain kinds of illness in infancy, and congenital syphilis operate as environmental causes of mental deficiency. Little has been done in the way of quantitative studies of causes to determine their relative importance, their occurrence, or the possibilities of treatment. The work has largely consisted of a small number of case studies. The quantitative studies have been based upon too small a number of cases to give them general validity. A combination of case and statistical study offers an opportunity for fruitful research which should have important practical bearings.

13. THE INTERRELATIONSHIPS AMONG SOCIAL PROBLEMS

From the foregoing outline of social problems it is obvious that social situations requiring public or private attention are intricately bound together. The effort to give every child a minimum of education brings with it the complicating conditions of personality maladjustment, mental deficiency, dependency, delinquency, and crime. Old age is not simply accumulation of years of life; it becomes a problem that is complicated by unemployability, illness, dependency, and insanity. Delinquency is complicated by poverty, inefficient parents, mental deficiency, and personality maladjustment. These illustrations emphasize the fact that social problems

have both social and physical causes and that treatment involves consideration of both factors. The social worker never has a case that can be treated as one simple problem. The social history of the case and the social milieu of the individual have to be taken into consideration. Social statistics is one of the methods for determining causes, points of concentration, and effectiveness of administration.

In cities it has been noted that several kinds of social problems concentrate in the same areas. Delinquency, crime, and dependency have been found high in the same census tracts in Indianapolis, and in Cleveland several kinds of diseases have been found to concentrate in the same census tracts. In Chicago poverty, crime, and delinquency are associated. The ecological study of social problems as suggested by the facts in these cities offers an interesting field for statistical research.

What this brief survey of the field of social problems has intended to point out to the student is the growing reliance upon statistical methods and the obvious need for more adequate statistical records and more systematic and continuous study of social problems as a public necessity. Social statistics is of primary importance to scientific work and to efficient administration.

CHAPTER II

Sources of Published Statistics

For thousands of years some kinds of social statistics have been kept by rulers and public officials. The clay tablets of ancient Babylonia reveal the fact that Hammurabi had a considerable amount of information about his people, particularly about the number and whereabouts of laborers and imperial slaves. The round population figures of the Old Testament indicate that the rulers had some conception of the numbers of their people and of their military man power. The Romans made estimates of the population of Rome and other cities, even if they did not take a careful census. The vast administrative system of the later Roman Empire necessitated some statistics. In the Middle Ages rather complete records were made of the population and status of the inhabitants of manors and feudatories. But it was not until the eighteenth century that social statistics in the modern sense began to be kept. Sweden has the longest record of population data of any country in the West. When the first census was taken in the United States, in 1790, one of the newest practices of modern governments was introduced. This census was a bare enumeration of the population for the purpose of determining representation in Congress. It had no scientific purpose, and its administrative purpose was limited to the relation between population and the number of representatives. In the decades since that date, the number of facts sought by the census-takers has gradually increased, so that it is now the most important and the most complete collection of sociological data in the country—this in spite of the fact that numerous agencies have arisen to collect social statistics. The census is our oldest effort at statistical sociology, and it is used by many students for a great variety of purposes. But there are other agencies which collect social data of great importance, and it is the purpose of this chapter to indicate the nature of the work done by some of these agencies. No attempt is made at a complete list of such

sources of social statistics. Aside from the description of statistics collected by the federal government, the sources mentioned merely illustrate types of agencies collecting statistics for administrative and scientific purposes.

I. THE VALUE OF A KNOWLEDGE OF SOURCES

A knowledge of the sources of statistical data is useful in several ways. It prevents needless duplication of work and waste of time. A social statistician who failed to acquaint himself with these sources would be like a historian who studied the history of the American Revolution without examining the collections of documents in the Library of Congress, the New York Public Library, and the Boston Public Library. If certain desired statistics are already in existence, the work of the investigator is lessened just that much, for he can get the published records and proceed with his study from that point. Another value of a knowledge of, and familiarity with, sources lies in the fact that it develops the habit of thinking of problems in terms of facts. Collections of data, like the census of population, may be presented in tables which can be further analyzed in the study of particular problems; or they may be presented with a complete analysis, as, for example, the monographs of the National Bureau of Economic Research. In either case, while familiarizing himself with them, the student is learning to ask for conclusions based upon facts rather than upon speculative reasoning. A third reason for knowing the most important sources of social statistics is that it develops the expectation that as time passes generalizations about social matters will be checked, rechecked, and refined by the constant appeal to facts. The natural sciences have made progress by innumerable accretions, some small and some large, for what has been discovered by one worker is published to the world of his fellow workers, and the body of the science grows. The social sciences will in the same manner progress from mere philosophy to something approaching science. Available social statistics constitute a part of the working tools of the social scientist, and are a part of the basis of action for the social administrator.

In social research an acquaintance with the sources of existing statistical data is indispensable. Two illustrations will make this clear: one concerns the costs of social institutions or organizations, and the other, the changing number of persons aided or affected in some way by these institutions and organizations. For several

decades the amount of money spent each year for the maintenance of public charitable and correctional institutions has apparently been increasing at a rapid rate. In Indiana expenditures for this purpose increased from $1,991,005.27 in 1910 to $5,145,640.55 in 1929—an increase in nineteen years of 158 per cent. But the purchasing power of money was changing during that time, and these dollars are not comparable, because in 1910 a dollar would buy more of the elements of maintenance than it would in 1929. When the dollars for the two different years are made comparable by the use of an index of general prices expressed in equivalent dollars, the amounts would be $2,073,964 and $2,874,659 respectively, or an increase of only 38.9 per cent in the money cost of maintenance. Putting it another way, after the price adjustment is made, the per capita expenditures for maintenance of state institutions in Indiana was 77 cents in 1910 and 89 cents in 1929. Obviously, one who is dealing with comparative costs of maintaining social institutions over a period of years should know something about indexes of the general price level.

The number of inmates in these institutions was 10,587 on the last day of the fiscal year in 1910, but it was 17,477 in 1929, an increase of 65 per cent. But as it stands, does this represent a measure either of the increase of the magnitude of social problems in Indiana or of an increased public interest in the persons for whom the institutions exist? Clearly it does not, because the population of the state has increased during these nineteen years. An accurate estimate of increase either in the problem individuals or in public interest must be based upon the number of persons in the state institutions per 100,000 population. In 1910 this was about 392; in 1929 it was about 539. Although this shows an increase in the relative numbers in the institutions, it is much less than 65 per cent. The social statistician can hardly begin the study of any problem that will not require reference to the census of population for standardization purposes. If he deals with money, he must have recourse to a general price index. Some problems will require still other information which may be available in published statistics. The student should have some knowledge of these sources and should know where to find the supplementary data he requires.

Teachers of the social sciences in high school and college find published statistics useful in their work. If a teacher knows his social statistics, his discussions of social problems or of general sociology will not have to be limited to qualitative analysis, but

can be supported by statistics. At best, the teaching of these subjects will be heavily weighted with opinion and speculation, but the more facts he knows and the greater his insight into their significance, the smaller the margin of opinion becomes. He is more independent of "authorities," and he acquires sufficient knowledge to be entitled to his own judgment in his field. By constant recourse to statistics applying to his subject, he develops the habit of appealing to facts. In other words, he teaches a body of knowledge and method, and not simply opinions.

Social administration is increasingly becoming the work of technically trained persons. The high executive may be less of a specialist than some others in his organization, but he depends for effective administration upon the work of experts. Writing of public administration, Leonard D. White says: "As these trends move on from decade to decade, they emphasize the decline of the amateur and the dominance of the expert. The amateur administrator long ago lost his hold on the national services and is disappearing in the larger local services as well."[1] This applies to private social organizations as well as to public institutions, bureaus, and departments. The administrator must know the facts which are important in his own organization as well as any others which have a bearing upon his efficiency. An understanding of statistical sources and studies in his field enables him to support, if not to replace, "hunches" by facts and by conclusions based upon a sufficiently large number of pertinent facts to make them trustworthy. It puts him in touch with the latest and best knowledge about problems similar to his own, and he benefits from the experience of others. He discovers new methods of analyzing his own problems and learns of sources of statistical data which aid in their study.

Two other values of a knowledge of sources should also be mentioned. Such knowledge provides a background upon which the investigator can block out the general situation in which he is interested. From these he obtains a picture of his problem and develops a perspective, both of which are important in planning studies and in interpreting the results of investigation. The other value lies in the fact that the investigator is enabled to relate his facts, which may involve special local and temporary variations, to the general trends shown by similar facts assembled from many

[1] White, Leonard D., "Public Administration," in the *Encyclopaedia of the Social Sciences*, Vol. I, p. 448. New York: The Macmillan Co., 1930.

sources. This kind of comparative study prevents hasty conclusions and the hasty adoption of policies.

2. FEDERAL GOVERNMENT STATISTICS

The United States government is the largest collector of statistics in the country. The public has little conception of the amount of statistical work done by the government, a situation partly explained by the fact that statistics do not make easy reading, and that it takes more than an ordinary newspaper reporter to write them up in a manner to make them front-page news. The statistical information which does find its way into news reports usually has some popular aspect that can be seized upon and played up. But the statistics which the student and the administrator find interesting are bound in bulky volumes or issued in paper-back bulletins. Dr. Lawrence F. Schmeckebier has performed a useful service in bringing together in one volume a description of the statistical work of our government.[2] Although much of the government's statistical work is concerned with matters not conventionally included in the category of social statistics, a brief outline of the types of this work will not be out of place here. The headings of Dr. Schmeckebier's chapters will suggest the scope of this work: (1) population in general, method of collecting data and the classification of the population; (2) special statistics of Negroes, Indians, Chinese, and Japanese; (3) dependents, defectives, and delinquents; (4) immigrants and emigrants; (5) occupations; (7) births; (8) deaths, diseases, and accidents; (9) marriage and divorce; (10) religious bodies; (11) education; (12) labor and wages; (13) women and children; (14) general agricultural conditions; (15) production of crops; (16) livestock; (17) livestock products; (18) production of minerals; (19) products of fisheries; (20) production of manufactured articles; (21) surveys of industries; (22) imports and exports; (23) land transportation and communication; (24) shipping; (25) domestic commerce; (26) water power and electric power; (27) prices; (28) finances of the national government; (29) public finances other than national; (30) general statistics of cities; (31) money and banking; (32) income and national wealth; (33) statistics of noncontiguous ter-

[2] Schmeckebier, Lawrence F., *The Statistical Work of the National Government.* Baltimore: Johns Hopkins Press, 1925. This is a publication of the Institute for Government Research.

See also Fry, C. L., "Making Use of Census Data." *Journal American Statis-Association,* pp. 129-138, June, 1930.

ritory, that is, Alaska, Hawaii, Porto Rico, etc.; (34) statistics of foreign countries; and finally (35) miscellaneous kinds of statistics. Published statistics on these subjects can be obtained either from the department or bureau issuing them or from the Superintendent of Documents, Washington, D. C.; in many cases they are distributed free, in others there is a small charge. "At the present time the statistical work of the United States government compares favorably, both in extent and quality, with that of any government in the world," says Dr. Schmeckebier. "In the field of manufactures, especially, there is nothing in the work of foreign governments that can be compared with our biennial statistics.[3] The truth of this judgment is further borne out by the fact that the Superintendent of Documents publishes the *Monthly Catalogue of Public Documents* so that anyone may examine the current publications in his special field. This catalogue is published like a journal, and the subscription price is fifty cents a year. It lists many documents which are not statistical, but the proportion of listed documents containing statistics is large.

The Bureau of the Census collects and publishes more statistics than probably any other division of the national government. Some of the information it collects will doubtless surprise the student who is familiar with the census mainly as the population of the nation, states, counties, and cities. The Bureau summarized the scope of its work recently in the following paragraphs:[4]

"The Bureau of the Census takes the decennial census of the United States covering population, agriculture, irrigation, drainage, manufactures, mines and quarries, distribution, and unemployment, and is continuously engaged in the compilation of other statistics covering a wide range of subjects.

"Statistics regarding the dependent, defective, and delinquent classes in institutions; public debt, national wealth and taxation; religious bodies or churches; and transportation by water are compiled every tenth year in the period intervening between the decennial censuses; and statistics of electric light and power plants, electric railways, telephones, and telegraphs every fifth year.

"A special census of agriculture is taken every fifth year following the decennial census; and a census of manufactures is taken biennially.

[3] *Op. cit.*, p. 1.

[4] *List of Publications of the Department of Commerce*, edition of May 15, 1930, p. 13.

"Statistics of births, deaths, marriages, and divorces are compiled annually; also financial statistics of cities and States; and statistics of prisoners in State prisons and reformatories, and of patients in hospitals for mental diseases and in institutions for epileptics and feeble-minded.

"At monthly intervals statistics are published relating to cotton supply, consumption, and distribution; to cottonseed and its products; and at approximately semi-monthly intervals during the ginning season reports are issued showing the amounts of cotton ginned to specified dates.

"The Bureau also collects monthly or quarterly data regarding the production or supply of many other commodities, including hides, skins, leather and leather goods, clothing, and wool. Current reports for these industries and commodities are multigraphed and issued as soon as the returns are tabulated. These reports are distributed free of charge, and a complete list of those available may be obtained from the Director of the Census.

"The Bureau publishes the monthly Survey of Current Business, compiling from various sources data regarding the movement of prices, stocks on hand, production, etc., for various lines of trade and industry, together with such other available data as may throw light upon the business situation."

The best known, and one of the most important, divisions of the work of the Bureau is the census of population. The headings of the schedule used for taking the census in 1920 were substantially as follows:[5]

Place of abode:

1. Street, avenue, road, etc.
2. House number of farm.
3. Number of dwelling house in order of visitation.
4. Number of family in order of visitation.
5. Name of each person whose place of abode was in this family.

Relation:

6. Relationships of persons enumerated to head of the family.

Tenure:

7. Home owned or rented.
8. If owned, free or mortgaged.

Personal description:

9. Sex.

[5] Schmeckebier, *op. cit.*, p. 18.

10. Color or race.
11. Age at last birthday.
12. Single, married, widowed, or divorced.

Citizenship:

13. Year of immigration to the United States.
14. Naturalized or alien.
15. If naturalized, year of naturalization.

Education:

16. Attended school any time since September 1, 1919.
17. Whether able to read.
18. Whether able to write.

Nativity and mother tongue:

Person enumerated:

19. Place of birth.
20. Mother tongue.

Father of person enumerated:

21. Place of birth.
22. Mother tongue.

Mother of person enumerated:

23. Place of birth.
24. Mother tongue.

Ability to speak English:

25. Is person enumerated able to speak English?

Occupation:

26. Trade, profession, or particular kind of work done.
27. Industry, business, or establishment in which at work.
28. Employer, salary or wage worker, or working on own account.

In addition to this information, the census of 1930 included several questions about unemployment. The unemployment schedule was filled out only by persons who usually worked, were then out of work, were able to work, and were looking for work.

The census of population is a complete enumeration of everyone in the country, and because of this fact it is invaluable as an aid to testing the representativeness of data collected for special studies involving population. For example, in a study of crime the offenders may be classified by age. Is there a concentration at certain ages? This question can be answered by comparing the age distribution of the population, as reported by the census, for the same area from which the crime data are drawn, with the age

distribution of the offenders. In many other ways the census of population may be used as a tabulation of the standard distribution of population characteristics, with which data collected for special purposes may be compared for testing the representativeness of the sample and for determining deviations from the normal distribution of the characteristics of the total population.

The unemployment census of 1930 undertook to enumerate all persons who were out of work because of the depression. The questions were so framed that they would exclude those who were idle because of illness, who quit their work voluntarily, who had been discharged for cause, who did not want to work, or who were out because of a seasonal decline in their occupations. Such a census had never been attempted before by the Bureau, and great difficulties were encountered in preparing the unemployment schedule. There was much criticism of the reliability of the results both before the census was taken and later when the results began to be published. The crucial problem was to define an unemployed person in such a way that the enumerator could recognize one and record the information asked for with a high degree of accuracy. This problem arises for the Bureau every time a decision is made to include an additional item in the census schedule.

Another important report of the Bureau deals with occupations, on which data have been obtained at each census since 1830. Prior to 1910 occupations were returned in terms of the industry with which the individual was connected. This was unsatisfactory because the types of occupations had changed greatly and because it did not permit the detailed analysis of occupations which later statistical inquiries necessitated. Consequently, the method of taking this census was entirely revised in 1910, and occupations were defined in terms of the worker and the particular job he did, regardless of the major industry of which he was a part. In the study of any problem touching child labor, school attendance, changing types of occupations, number engaged in gainful occupations, and geographical distribution of occupations, the census data are of inestimable value.

The Bureau of the Census has been collecting various kinds of institutional statistics since 1830, the first being those on the blind and deaf obtained in that year. In 1840 data were collected on the insane and feeble-minded, and in the census of 1850 data on paupers and delinquents were included for the first time. Four special reports were published in 1904, 1910, 1915, and 1923,

giving data concerning benevolent institutions; this classification includes children's homes, day nurseries, hospitals, dispensaries, permanent homes, temporary homes, and schools and homes for the blind and deaf. Since 1890 all Bureau reports concerning dependents, defectives, and delinquents have been issued as special reports and not as parts of the decennial census. An annual census of prisoners, the insane, the feeble-minded, and the epileptic in institutions has been taken since 1926. These reports are of great value in showing growth but for the most part they reflect only the magnitude of the problems as indicated by institutional populations, capital investments, and cost of maintenance. They do not purport to measure the extent of dependency, defectiveness, and delinquency in the whole population; but, used with a full consciousness of their limitations, these census reports are valuable.[6]

Marriage and divorce have been the subject of census publications since 1899. In that year a compilation of marriages and divorces from 1867 to 1886 was made and published by the Bureau of Labor. Later another report was issued by the Bureau of the Census which included the data of the older report and brought them down to 1906. It was expected that a further report would cover the period from 1907 to 1916, but the war intervened, and this report was limited to marriages and divorces for the year 1916. Beginning with 1922, however, the Bureau of the Census has published annual statistics of marriage and divorce. The data are given by geographical divisions as follows: the nation, groups of states, states, and counties. The number and rates of marriages for the population 15 years of age or older are given, and divorces are presented in tables showing the number by age and sex, the cause of divorce, children involved, and rates per 1,000 married people. The collection of these statistics is not complete for the entire country. "The statistics of marriages," says a bulletin of the Bureau, "are now obtained from some office of the State government in 29 States, and the statistics of divorces are likewise obtained from State officials in 16 States. In the other States county officials furnish the information."[7] The reliability of the reports of the county officials is questionable, and, in addition, all of them do not report regularly. But the reports for the 29 states on marriages and the 16 states on divorces are probably fairly representative of the country as a whole. In states where new laws

[6] Schmeckebier, *op. cit.*, Chap. V.

[7] *Annual Report of the Director of the Census*, June 30, 1930, p. 20.

affecting marriage and divorce have been enacted such statistics enable students and administrators to determine to some extent the effects of this legislation.

Vital statistics are valuable not only in themselves but for their use in the study of a variety of social problems. Since 1915 the Bureau of the Census has annually published the statistics of births in the registration area. This area included the District of Columbia and the states which provided by law for the registration of births. In 1930, the registration area covered 46 states, South Dakota and Texas being the only ones not included; in South Dakota, however, one city reports, and in Texas eight cities report. Statistics of deaths are now available for the same states. The reporting of deaths according to a registration area began in 1880 with Massachusetts and New Jersey; but by 1890, six states were reporting systematically, and this number has steadily increased since that time. Since the organization of the permanent Bureau of the Census in 1900, annual statistics of deaths have been published as "Mortality Statistics." Mortality and birth rates are estimated each year, but only in the census years can an accurate calculation be made.

A census of religious bodies has been made since 1850. Questions to obtain this information were asked at the time of the regular census in 1850, 1860, 1870, and 1890. No general statistics were collected in 1880, though a report on cities did give certain information for the cities only. The law was changed after 1890 to provide that a census of religious bodies should be made every ten years, but not in the year of the decennial census, and in accordance with this provision special reports were prepared in 1906, 1916, and 1926, giving the number of organizations of each denomination, the number of communicants by denominations, and the geographical distribution. These reports are of great usefulness for studying the church as a social institution.[8]

In addition to the volumes of data published by the Bureau of the Census, a number of important monographic studies of special problems have been made in recent years. These monographs deal with certain of the more important subjects covered by data collected by the Bureau, and may be obtained from the Bureau or from the Superintendent of Documents.

The Children's Bureau of the Department of Labor is engaged in the work of promoting the welfare of children. Dr. Schmecke-

[8] Schmeckebier, *op. cit.*, Chap. XI.

bier says of this Bureau: "The Children's Bureau of the Department of Labor is concerned with the study of questions relating to child life, and a portion of its work has a statistical basis, the remainder being descriptive and expository and dealing with such subjects as child-labor laws, illegitimacy, and health of mothers and children. The statistical publications of this Bureau are not issued at regular intervals, and do not form a series covering the same field for a number of years. Each one relates to a specific topic in a limited area, and is a complete study for the particular period, topic, and area covered."[9] What Dr. Schmeckebier says is correct, but it does not lessen the value of the publications of the Children's Bureau for their own purposes or for study by those who wish to gain an understanding of how studies in child welfare are made.

Since the publication of Schmeckebier's work this Bureau has begun the publication of a regular series of statistics, known as the monthly reports of the Registration of Social Statistics. It was begun July 1, 1930, and at present is issued in monthly reports. The Bureau provides a schedule which is mailed to a number of large cities which have agreed to coöperate with it in collecting the data. The schedule calls for information regarding family welfare and relief, mothers' and widows' pensions, non-institutional service to ex-soldiers and their families, free legal aid, travelers' aid, dependent or neglected children in foster homes or in institutions, applications for the care of dependent or neglected children, case work for such children, children in detention homes, protective case work for young people, care of children in day institutions, adult probation, temporary shelter for homeless or transient persons, maternity homes, hospital in-patient service, clinic and dispensary out-patient service, medical and psychiatric social service, public health nursing, and school health service. Each classification represents a table which shows a detailed analysis of data pertaining to it. Reports from a city are not used unless they include information from substantially all the agencies rendering the particular service. In this respect the Registration of Social Statistics differs from all other plans of social reporting. The public agencies print reports of their own statistics, but they omit reports of similar work by privately supported agencies. In some cities a community chest obtains statistical reports from all its member agencies, but these reports leave out the publicly sup-

[9] *Op. cit.*, p. 176.

ported agencies. The reports of the Children's Bureau attempt to cover completely the cities for which such data are published. They began much as the Bureau of the Census did in the case of vital statistics, namely, with a registration area.

The amount of research that can be done on the basis of the Children's Bureau reports is as yet quite limited, but suggestive analyses may be made. This work was taken over from the Joint Committee of the Association of Community Chests and Councils and the Local Community Research Committee of the University of Chicago, and some analysis of the data has been published in the form of reports for 1928 and 1929.[10] The Joint Committee got the work of reporting under way, and then it was taken over by the Children's Bureau. These two annual reports and the current monthly reports of the Children's Bureau are exceedingly useful in teaching social statistics. They offer opportunities for the analysis of the social statistics themselves, and these data can be the basis for formulating certain problems which require the use of the census of population and of occupations. Vital statistics may also be introduced and correlated with the social data. Thus, there might be set up a research project which would extend over a considerable period of time and would involve the use of a good many of the common statistical methods.

Another governmental bureau whose statistical reports may not be classified wholly as social statistics but much of which are social statistics is the Bureau of Labor Statistics. "In carrying out the purpose for which the Bureau of Labor Statistics was created," says a bulletin of the Bureau, "data are collected in various ways from various sources—by personal visits of agents in the field and from correspondence, by consulting reports, trade journals, and other publications, by contract with experts to make special studies, and in other ways. All of the material in the publications of the bureau, whether prepared in the bureau or contributed by persons specially contracted with, is carefully edited in the office, and all facts and figures verified, whenever practicable, by comparison with the original sources."[11] Some of the statistics published by the Bureau are primarily economic, but for the most part they have much wider social implications.

[10] See McMillen, A. W., *Measurement in Social Work,* for a study of the data for 1928 and 1929. Chicago: University of Chicago Press, 1930.

[11] *Methods of Procuring and Computing Statistical Information of the Bureau of Labor Statistics, Bulletin No. 326, 1923,* p. 1. Much of the material contained in the following paragraphs is taken from this bulletin.

The *Monthly Labor Review*, the official periodical of the Bureau, has been published since 1915 and is the Bureau's medium for the presentation of reliable information concerning labor in all its aspects. The following subjects are given special attention in the *Review*: "Wholesale and retail prices and cost of living; wages and hours of labor; productivity and efficiency of labor; minimum wage; industrial relations and labor conditions; woman and child labor; labor agreements, awards, and decisions; employment and unemployment; vocational education; housing; industrial accidents and hygiene; workmen's compensation and social insurance; labor legislation; decisions of courts relating to labor; labor organizations; strikes and lockouts; conciliation and arbitration; immigration; coöperation; employees' representation; welfare work; profit sharing; etc."[12] Several of these subjects obviously will not receive statistical treatment in the *Review*, but many of them cannot be discussed without recourse to statistics, and most of them involve statistical considerations at some point. Articles which present statistical tables deal with the following subjects: changes in membership in unions connected with the construction industry, transportation unions, mining, oil and lumber unions, paper, printing and bookbinding unions, clothing unions, etc., between 1926 and 1929; the development of credit unions by states and cities; unemployment surveys in several cities; industrial accidents; consumers' coöperation; labor turnover; industrial disputes; housing; wages and hours of labor; the trend of employment; wholesale and retail prices; the cost of living.[13] For the student of labor problems and for the research worker there is much material of value in this one number of the *Review*. Other numbers cover similar ground but with a varying amount of statistical material on different topics.

No other source is so comprehensive in its treatment of wages, hours of labor, pay-roll data, and labor turnover as are the *Monthly Labor Review* and certain special reports of the Bureau. For example, here is found authoritative information about increases or decreases in wage rates. Prominent politicians have been known to assure the public, in the depression of 1929-—, that wage rates were being maintained, when the monthly reports of the Bureau showed a strong tendency for employers to reduce them. Changes in hours of labor are recorded at length. The

[12] *Op. cit.*, p. 52.

[13] *Monthly Labor Review*, February, 1930, Vol. 30, No. 2.

indexes of factory and of steam railroad employment show the national tendency toward depression or prosperity, or they reflect the displacement of men by machines. A decline in the amount of pay-rolls reflects declining employment or reduction of wages. Such data and indexes have a wide range of uses in the study of social problems.

Closely allied to wage rates and pay-rolls, and also related to wholesale and retail prices, is the cost of living. The index of the cost of living published by the Bureau of Labor Statistics is constructed on the basis of several hundred items entering into the family budget. The year 1913 is taken as 100 per cent, and indexes of subsequent years are expressed as percentages of the average cost of living in that year. If wage rates have increased, have they gone up as fast or faster than the cost of living? The index of the cost of living makes this kind of comparison possible. Prices of some articles of consumption change more rapidly than others. In order to show which items in the family budget are lowering or raising the cost of living, index numbers of the cost of separate items—food, clothing, rent, fuel and light, etc.—have been computed and are published currently. Index numbers of retail and wholesale prices are also published. While these are similar to the index for the cost of living, they are not identical, since the items entering into the latter index are weighted in accordance with their importance in the family budget.

Industrial accidents play an important part in many social problems beyond the injury to the worker and his temporary loss of wages. Workmen's compensation laws have helped to relieve the immediate economic distress of the worker's family, but they only help. Permanent partial disability leaves him with less earning power, and total disability removes him entirely as a source of income to his dependents. Dependency of his family may result with a long array of attendant evils. The Bureau of Labor Statistics publishes a summary of statistics of industrial accidents which are gathered by the National Safety Council (Chicago); it also shows comparative rates for different industries and for the same industry in different years. A comprehensive study of dependency and charitable relief has to take into account the permanent effects of industrial accidents, and this fact alone gives the Bureau reports added importance for students of social statistics.

The United States Public Health Service is another source of important statistics. Ill health has many ramifications which ap-

pear as complicating factors in a variety of social problems. The collection of morbidity statistics, however, has lagged far behind the development of statistics of births and deaths. It is a simple matter to record a birth, and it is equally simple to record a death, though more difficulty may be encountered in stating the cause of death. When the whole range of morbidity is considered, it is not surprising that statistical reporting lags. More advance has been made in reporting communicable diseases than others. Such diseases as diphtheria, measles, smallpox, and tuberculosis are more easily diagnosed than some others, and the public has a very real interest in knowing the time and location of cases. But some communicable diseases, like gonorrhea and syphilis, are not adequately reported, because most physicians still regard such information about their private patients as confidential and refuse to report it unless compelled by law. Cases of venereal diseases in public hospitals are likely to be reported, but this group constitutes a small proportion of all such cases. Nevertheless, some progress is being made in reporting morbidity. The United States Public Health Service, through its Division of Sanitary Reports and Statistics, is attempting to systematize the national reporting of various diseases, particularly communicable diseases. Since this service is organized all over the world, special precautions may be taken if yellow fever, cholera, or other similar diseases appear in a foreign port with which Americans have frequent contact. "The collection and dissemination of information concerning the prevalence of disease is of increasing importance in this age of speedy transportation facilities. For instance, it is possible that a person infected with typhoid fever may, even by motor, traverse the entire width of the country before completion of the incubation period of this disease."[14] In order to acquaint the public with the location and prevalence of reportable diseases, the Public Health Service publishes a weekly report for general circulation. This is of primary importance to public health officials and to others associated in some way with public health work, but the data published are frequently useful to research workers in other fields who require health data for problems under investigation.

As stated, the Public Health Service extends throughout the world. ". . . every consul and consular officer stationed abroad makes a weekly report to the Public Health Service as a part of

[14] *Public Health Reports,* Vol. 46, No. 6, p. 285. These reports are issued weekly by the United States Public Health Service.

his routine duties. The reports are made on forms provided by the Public Health Service and bearing a list of the more important communicable diseases. The consular officer obtains reports from health officials of the country to which he is accredited, and from these reports and such other sources as are available he fills in the information required on the form and mails it to the Public Health Service. These reports by mail cover the following diseases: Cerebrospinal meningitis (epidemic); cholera, Asiatic; cholera nostras, cholerine, or gastroenteritis; diphtheria; measles; plague, human; plague, rodent; poliomyelitis (acute anterior poliomyelitis or infantile paralysis); scarlet fever; smallpox; tuberculosis; typhoid fever (enteric fever, typhus abdominalis); typhus fever (typhus exanthematicus); and yellow fever."[15] "In the domestic field the Public Health Service is kept informed of conditions by weekly reports mailed in from local health officials in 570 cities of 10,000 or more population. These reports cover the prevalence for their respective territories of the following diseases: Chicken pox, diphtheria (carriers not included), influenza, measles, mumps, pneumonia (all forms), scarlet fever, smallpox, tuberculosis (all forms), typhoid fever, whooping cough, cerebrospinal fever, dengue, lethargic encephalitis, pellagra, poliomyelitis (infantile paralysis), rabies (in man) (developed cases), rabies (in animals), typhus fever."[16] The second half of the weekly reports gives statistics of these diseases in two parts: first, the United States, and, second, foreign nations and the island possessions of the United States.

Besides the weekly reports of morbidity, the Public Health Service makes special surveys of public health work, one of the most recent of which was a study of this work in Oklahoma.[17] This report covers a study of the law creating the State Board of Health, the administration of the department, the organized medical profession in Oklahoma, the state educational authorities, and unofficial health agencies; and it mentions two outstanding defects: "1. The failure to do any more than scratch the surface in the most important field of public health, viz., the hygiene of the preschool child. 2. The lack of properly organized local health units to apply, locally, the policies of the State Health Department."[18] The report is largely nonstatistical, but it is based upon

[15] *Ibid.*

[16] *Ibid.*, p. 286.

[17] *Public Health Reports*, March 13, 1931, Vol. 46, No. 11, pp. 575-598.

[18] *Ibid.*, p. 577.

the collection of facts which it seemed unnecessary to present in detail.

Other departments and bureaus of the national government publish statistics useful to the social statistician, but the work of those described above constitutes the most important sources. In special research work economic and technical statistics may be desirable, and these can be obtained by applying to the proper department or bureau. If it is not known what office publishes them, recourse may be had to Dr. Schmeckebier's book, *The Statistical Work of the National Government*, which gives a brief description of all types of statistical work done and states where the various reports may be obtained.

3. SOCIAL STATISTICS OF STATES

Many statistical reports are issued by departments of the state governments. Those of most interest as social statistics are the annual, biennial, and quadrennial reports of state boards of administration, control, charities and corrections, and public or social welfare. The Russell Sage Foundation[19] has listed forty states and the District of Columbia as having some kind of boards which concern themselves with what is commonly called public welfare work. Some of these boards merely supervise the financial affairs of state institutions, while others collect statistics and act in an advisory relation to these institutions, and still others are the central administrative body for them. All these boards publish statistical reports. For some purposes these reports are more useful in teaching research methods than the summaries covering similar subjects published by the national government, because the facts are given in greater detail. We shall give a brief description of the social statistics published by a few of these state departments.

One of the oldest is the Department of Public Welfare of Massachusetts. It was organized in 1863 as the Board of State Charities, and its long history makes its annual reports of great value in the study of the development of public welfare work in Massachusetts. The report is divided into sections for aid and relief, child guardianship, and juvenile training, under the direction of which come such services as indoor and outdoor poor relief, mothers' aid, care of handicapped children, of dependent and neglected children, and of delinquent children in institutions. An

[19] *Directory of State Boards,* bulletin of the Russell Sage Foundation Library, No. 96, August, 1929.

annual report made of these various kinds of work shows both the number given service and the cost of the service to the state. The most adequate table is that of statistics of public poor relief.[20] The Department also supervises private charitable corporations and publishes statistics of the volume and cost of their work. Statistics of adult delinquents and criminals and of insanity and mental deficiency are not published by this Department, but by special commissions created for these types of work.

The Indiana Board of State Charities was organized in 1889. At the time of its organization it was modeled to a considerable extent upon the corresponding department in Massachusetts, but it differs in some important respects. It has advisory authority only in so far as the conduct of public welfare work is concerned, with one exception: that boarding homes for children are licensed and subject to inspection by officers of the Board. Its only direct social work is done in placing and supervising dependent and neglected children in free homes. Its major function probably is the collection of statistics from institutions and agencies, which are required by law to report to it at stated times concerning the work of the quarter or the year. The Board publishes a monthly bulletin, a large part of which consists of statistics, and the number of the bulletin which gives the annual report is almost entirely statistical. Statistics of crime and delinquency, mental disease and deficiency, dependent and neglected children, county general and tuberculosis hospitals, and indoor and outdoor poor relief appear in this number. Comparative statistics for a number of years are usually given, and occasionally the annual report gives statistics by years as far back as they have been reported to the Board. The tables are well-prepared and intelligible. The one giving the reports of the county poor asylums distributes the population in these institutions by sex under the following headings: feeble-minded, insane, epileptic, paralytic and crippled, deaf, blind, senile, sick, able-bodied, total population at the end of the fiscal year, and total admissions during the year by counties. This kind of an analysis makes the report particularly useful for student statistical analysis, because it discriminates the different types of persons requiring indoor poor relief. Another table gives the distribution by age and sex. The table presenting statistics of outdoor poor relief gives comparative data back to 1890 for some of the following headings: number of

[20] *Annual Report*, Department of Public Welfare of Massachusetts, November 30, 1928.

families and of single persons; number of males and females; and the amount of outdoor relief each year for the state as a whole.

The Illinois Department of Public Welfare, organized in 1917, publishes an annual report containing elaborate statistics of institutions and agencies dealing with the insane, the criminal, the delinquent, the dependent and neglected child, the handicapped child, the feeble-minded, and the poor. This report is of great value for studies in social statistics because of the detail presented, particularly in the case of statistics of crime and insanity. The statistics on crime are given in tables which show race and nationality, type of crime committed, age distribution, and educational attainments. Statistics of insanity show age, sex, county of residence, race or nationality, type of mental disorder, marital condition, religion, duration of hospital residence, rate per 100,000 population, and several other less important facts. The details concerning the insane are as complete as such things generally are in the report of a single hospital for mental disease.

The annual reports of state departments of public welfare can usually be obtained free upon request. They offer a wealth of laboratory material for classes in social statistics, and some of them, such as the Indiana analysis of people receiving poor relief and the Illinois analysis of those in prisons and hospitals for the insane, may be used for more extensive statistical research.

4. STATISTICS OF PRIVATE ORGANIZATIONS

During the last twenty-five years there have come into existence a large number of private organizations, a part or all of whose work is social research. Some of them collect general statistics dealing with a wide range of subjects, but most of them use their statistics as the data for particular research projects in which they are interested. The publications of these organizations are of interest to students of social statistics from two points of view: first, the completed research project adds something to the accumulating knowledge of social institutions and social organization by the application of scientific methods; second, the statistics published as such, and not as finished research, are useful for further analysis and for their relation to other series of statistics. A short account of the work of several of these organizations will be given for the purpose of indicating the type of work the student will find on examining their lists of publications.

One of the oldest and best known of these organizations is the

Russell Sage Foundation, incorporated in 1907. The Foundation has as its purpose "the improvement of social and living conditions in the United States of America." The means of achieving this purpose have been largely social research and the publication of the results.[21] One of the most recent publications of the Foundation is *A Bibliography of Social Surveys*, which lists upward of 2,700 reports of surveys published up to January 1, 1928. Students doing research can consult the classified lists in this book and find out where and when other studies similar to their own have been made. If this is done and the reports examined, this book will contribute measurably to the scholarly character of research bearing upon social work. Another book, *Employment Statistics for the United States*, edited by Ralph G. Hurlin and William A. Berridge, presents a definite plan for the collection of employment and pay-roll statistics and suggests uses for them. It was worked out by the Committee on Government Labor Statistics of the American Statistical Association and represents a thoroughly competent judgment on the methods of collecting employment statistics. Occasionally a monograph is published, such as *The Longshoremen* by Charles B. Barnes. This is a study of working conditions, with special reference to the effects of seasonal variations in the employment of longshoremen. It is not entirely statistical, but some parts of it are based upon statistics of employment and earnings of this group of workingmen. The *Social Workers' Guide to the Serial Publications of Representative Social Agencies* by Elsie M. Rushmore provides a check list of the publications of over 4,000 institutions and organizations, and is another very useful index of sources. Many other books and pamphlets have been published by the Foundation, but these will suffice to indicate the importance of its work for research workers and social administrators.

Besides the publication of occasional books and pamphlets, the Department of Statistics of the Foundation began the collection of monthly statistics from relief agencies in a number of large American cities; in February, 1932, 62 cities reported. Twenty-nine of these cities are represented by all the important relief-giving agencies; the reports for the others cover only a part of their agencies. These statistics are compiled monthly, published, and mailed to the coöperating agencies and other organizations and

[21] For a complete list of the publications of the Foundation see *A Catalogue of Publications*, issued by the Foundation in 1930.

individuals who have arranged to get them. This project was started in 1926, about two years before the Registration of Social Statistics which the Children's Bureau[22] is now conducting. The statistics collected by the Bureau cover all types of social agencies in the cities from which they get reports, whereas the Foundation is collecting only relief statistics, mainly those of private relief agencies, though some public agencies are included. The raw data thus collected has two uses: discovering trends in relief and calculating a seasonal index of relief. The material can be used to advantage for laboratory purposes in teaching social statistics.

The only organization engaged exclusively in population research, as opposed to the mere collection of population statistics, is the Scripps Foundation for Population Research operated in connection with Miami University. Its work is based largely upon statistics gathered from all parts of the world. Its purpose is not to publish statistics *per se*, although in its publications a good deal of statistical material is given in tables which may be useful to other students in their own work. Two examples will show the type of work done by the staff.[23] Mr. Whelpton's object, as the title of his article suggests, is to estimate the growth of the population of the United States for the next fifty years. Beginning with the population as shown by the census, he uses birth rates, death rates, immigration statistics, and those of rural-urban migration for his estimates. The results of these estimates are given in tables, and a full explanation of the method accompanies their description. Mr. Thompson is concerned with the effects of the changing rates of growth of national populations upon the control of the land area of the earth. He gives tables showing the age distribution, birth rates, death rates, and natural-increase rates of most of the principal nations of the earth, and points out that a struggle among nations for control of the land is likely to come because of the differential rates of population increase. Such an analysis and such data as are presented in this article may have a bearing upon many social problems now being studied.

The National Bureau of Economic Research is another organization whose purpose is research—statistical and otherwise, but particularly statistical. The Bureau has been especially interested in

[22] See p. 49.

[23] "Population of the United States, 1925 to 1975," by P. K. Whelpton, *American Journal of Sociology*, Vol. XXXIV, No. 2, pp. 253-270. "Population," by Warren S. Thompson, *American Journal of Sociology*, Vol. XXXIV, No. 6, pp. 959-975.

business cycles, the distribution of income in the population, and variations in employment. Two publications, *Trends in Philanthropy*, by Willford I. King, and *Corporation Contributions to Organized Community Welfare Services*, by Pierce Williams and Frederick E. Croxton, deal with problems of particular concern to social welfare agencies. Its publications dealing with variations in employment are especially important for the social administrator whose volume of work declines in times of high employment and rises during low employment. The income studies reveal the fact that 98 per cent of the population receives about 85 per cent of the annual income, whereas 2 per cent of the population receives about 15 per cent of the annual income. From the lowest income classes arise many of the social problems with which agencies have to deal.[24] Elaborate tables on income are given in Leven and King's book, as well as in other publications of the Bureau. The publications are useful as secondary statistical sources as well as for the conclusions derived from the analysis of data.

Another organization chiefly interested in carrying on research projects, but which has published some statistics as such, is the Institute of Social and Religious Research. It was organized after the collapse of the Interchurch World Movement to salvage the material collected by this organization, and it has gradually developed upon broader lines. Its research has to a considerable extent been concerned with rural social questions. Several of its publications deal with the rural church, and a few have been reports of surveys of urban churches. One, *Middletown*, by Robert S. and Helen M. Lynd, has received the widest attention; this is an experimental application of the methods of social anthropology in a case study of a small industrial city. In all the publications of the Institute statistical material has been used extensively, and some of it is presented in such form that it may be utilized in connection with studies by other workers. The publication most useful as a statistical resource is *American Villages*, by C. Luther Fry, which is composed of population statistics of small towns. These statistics are not available in any of the publications of the Bureau of the Census. A special tabulation of certain data in the files of the Census Bureau was necessary to get the material for this book, and for this reason it is a particularly important source for a certain kind of population data for the social statistician.

[24] *Income in the Various States*, by Maurice Leven and Willford I. King, 1925. See p. 291ff. for the above estimates.

There is frequent need of different kinds of index numbers in the study of social statistics, and the best source for all kinds of index numbers is probably the *Standard Statistical Bulletin* published monthly by the Standard Trade and Securities Service. The table of contents in the 1930-31 edition contains 326 classifications of economic statistics and economic indexes, under each of which are from 1 to 98 subdivisions. Standard republishes in its *Bulletin* the principal index numbers of production, sales, and prices which are made by all other economic statistical organizations. The social statistician is not often called upon to compute his own economic indexes. It may be necessary in carrying on a particular kind of research, but he can usually find a suitable index which has been computed by specialists. Therefore, a few of the indexes he is most likely to use will be presented and their uses suggested.

Probably more frequent use is made of price indexes than of any other type, for in many investigations the social statistician is concerned with costs over a period of time. Since the purchasing power of the dollar is continually changing, it is necessary to reduce actual expenditures to comparable dollars. For example, when a state institution or department or a federal department is seeking increased appropriations to carry on its work, it is important to show the legislators that what is asked for is partly to maintain the same standard of work in a period of rising prices that was formerly attained with less money. Conversely, when prices are falling, the legislature may insist upon reducing aggregate appropriations because as time passes the dollar has greater purchasing power. One of the price indexes published by Standard is the Index of General Prices constructed by the Federal Reserve Bank of New York. According to this index, $1.00 in 1913 would buy as much as $1.79 in 1929. Thus, if a state department received $1,000,000 in 1913, it would need $1,790,000 in 1929 to maintain the same standard of work, assuming the degree of efficiency to be the same at both periods. In 1929 a dollar bought only a little over half as much labor, wholesale and retail commodities, and rent as it did in 1913; its purchasing power had decreased markedly. After the onset of the 1929-— business depression prices began to decline more rapidly; that is, the purchasing power of the dollar rose. Legislatures in 1931 reduced many appropriations from the level of the preceding biennium, not, however, because prices were falling but because the public demanded reduced taxes. If prices have at present fallen sufficiently,

it may be that little or no retrenchment is necessary on the part of state institutions to keep their work up to the standard of the last two years.

An index of the cost of living is sometimes useful in social statistics, and the one most widely used is that published by the United States Bureau of Labor Statistics and republished by Standard. An index of the cost of living differs from an index of wholesale or retail prices, or the general price level. It is nearest to an index of retail prices, but usually the weighting scheme is different and that affects the index as finally determined. The cost-of-living index is computed from retail prices of goods bought by families for consumption, and the quantities consumed are weighted by the relative importance of the commodity in the family budget. It is useful to relief agencies, which intend to maintain a minimum standard of physical efficiency in their clients, when the size of an allowance is determined. In a number of cities the relief agencies have studied local retail prices and have made up their own budgets on the basis of the cost of living, but in most localities this has not been done. The index of the cost of living as computed by the Bureau of Labor Statistics might be used to advantage by small agencies and public poor relief administrators to determine the amount of money required to purchase food, clothing, shelter, fuel, etc., sufficient to maintain physical efficiency. Enlightened employers might even use it as one factor in setting the minimum wage level.

Indexes of employment, wages, and general production are of less obvious use to social administrators and statisticians than those of prices and of cost of living, but they may be needed occasionally. Wage indexes are important in all social problems touching on the standard of living. For instance, if wages go up at the same rate as the cost of living, the standard of living is being maintained but not improved; if the cost of living rises faster than wages, then the standard of living is being reduced. Because the volume of employment responds quickly to changes in production, employment and production indexes may be useful in many ways, though not so obviously as certain other indexes. Declining production is soon followed by declining employment, and declining employment results within two or three weeks in a rise in the number of families receiving charitable relief. Thus, the close study of the trend of production indexes would be a general guide to social agencies in looking ahead, if not in making definite plans.

Crimes against property can be forecast in some degree by the use of a general business index. When a depression is seen coming, it might be a good time for police departments to add a few men to the force and to cover more carefully the areas in which crime is most prevalent. Other uses could be found for such indexes.

5. STATISTICS OF INDIVIDUAL AGENCIES AND INSTITUTIONS

Most social agencies and institutions make annual reports of some kind, and some of them publish statistical material at irregular intervals. The annual reports of state institutions are likely to contain tables which may be of considerable value to students of social problems in their own states, because they generally give more detailed statistics than a central collecting agency publishes. This is notably true of the reports of hospitals and prisons in some states. The further removed a statistical report is from the agency or institution which made the original records, the less detail there will probably be in it; and because this is generally true, it is desirable for students to have access to some of the reports put out by the institution or agency which made the original records.

The Cleveland Health Council is an example of a social agency which publishes statistical studies at irregular intervals. The Council is the research and coördinating agency of all the principal health agencies in Cleveland. These agencies require many studies of public health matters. The Council has published studies of density and fluctuation of population in different parts of the city, distribution of inhabitants by country of birth, distribution of cases of influenza from December 1, 1928, to January 31, 1929, and of mumps from August 1, 1926, to July 31, 1927, distribution of families served by family case-work agencies in 1928, and other special studies.[25] The population returns of Cleveland, largely through the efforts of Mr. Howard Whipple Green, Director of Research and Secretary of the Health Council, have been tabulated by census tracts for 1910, 1920, and 1930, and the Health Council publishes these data with a street index. The Council also publishes an occasional supplement in intercensal years, giving estimates of the population by tracts. All this material has a variety of practical uses.

The Institute for Juvenile Research of Chicago has published

[25] For maps showing the above studies, see "Facts, Figures and Fiction in Social and Health Statistics," by Howard Whipple Green, in *New England Journal of Medicine,* Vol. 202, No. 16, April 17, 1930, pp. 771-778.

some highly important studies of the distribution of crime and delinquency in that city. The plan of these studies has to a large extent followed the theory of "human ecology," that is, the study of the distribution of groups of the population and the forces which caused these groups to segregate and to survive as they have. The most interesting of these studies from the point of view of the social statistician is *Delinquency Areas* by Clifford R. Shaw and his associates.[26] Maps show the utilization of the land covered by the city, and the residence of each delinquent is indicated by a dot on the maps. This device shows the points of concentration of delinquency, and makes clear the relation of delinquency areas to the railroads and industries. Delinquency and crime are thus shown to be heavily concentrated in the downtown sections of the city and around the stockyards, but the concentration decreases rapidly as one gets away from the Loop District. Thus conditions which favor the residence of criminals and delinquents are emphasized by this method of study. Students of crime statistics will find this and other studies of the Institute suggestive for their own work.

The Research Bureau of the New York Welfare Council at irregular intervals publishes statistics which are useful to students in New York and elsewhere. These statistics usually relate to specific social problems in Greater New York, but often they have an important bearing upon general problems concerning other cities as well. The Council has recently published *A Guide to Statistics of Social Welfare in New York City*, by Florence Dubois, listing hundreds of studies of social problems made in New York, and making a knowledge of this work easily accessible to anyone waiting to consult one or more of these studies.

Social Science Abstracts, a monthly journal of abstracted magazine articles on social science and social work, has in each issue a section on "Research Methods," and one of the divisions of this section is devoted to statistical methods. Many of the articles are statistical studies of social problems with which every social worker and social investigator is concerned. At the end of each year an index number of *Abstracts* is published, listing by subject and author all the abstracted articles appearing during the year. This journal should be consulted at an early stage in any study, to see what has already been done on the problem under consideration.

A number of the larger cities of this country have created city

[26] *Delinquency Areas*, Clifford R. Shaw, *et al.* University of Chicago Press,

census committees whose purpose is to get the census returns tabulated by tracts and then to publish this material in a form available for administrative and research uses. The census tract is a small area, varying in size in the same city and in different cities, laid out with a view to enclosing a small homogeneous population —one similar as to race, nationality, economic status, etc. The New York Census Committee, the first of these committees, published the returns of the 1910 and 1920 censuses by sanitary districts which were accepted as adequate census tracts. The volume containing the 1920 New York census data distributed by census tracts was as large as the population volume issued by the Bureau of the Census for the whole country. The census-tract plan makes possible highly dependable statistical studies of social and health problems in large cities, and 15 of our larger cities used this plan in 1930. The published volumes, which can usually be bought from the census committees in the respective cities, furnish a basis for a great deal of social research, and provide data that is of inestimable value for teaching purposes. Because of their detail, these volumes are more useful for teaching and for intensive research in particular cities than is the material published by the Bureau of the Census.

6. STATISTICAL ORGANIZATION

An examination of the history of any organization whose object is the collection or analysis, or both, of statistics will reveal a growth. The organization frequently was created for a specific purpose and became a general collecting agency, or it started out to do one thing and developed over a period of years so that it does many others not conceived in its initial stages. For example, as has been said, the United States census began as an enumeration of the population for the purpose of apportioning representatives in Congress. For 110 years the census organization was set up every ten years, and when the collection, tabulation, and publication of the returns were completed it was disbanded. However, additional items were gradually added to the enumeration schedule, and in 1900 the organization was made permanent and established as the Bureau of the Census. At first only decennial data were collected; now the Bureau collects several kinds of data which are published monthly or annually. The Institute of Social and Religious Research has extended its work and has included in its scope many studies not even thought of by the original

organization. However, allowing for the certainty of change in any statistical organization as time passes, some characteristics are common to most of them, and these may be set forth briefly.

Three kinds of statistical organization may be distinguished roughly by the purpose of collecting the statistics: (1) statistics as bookkeeping; (2) statistics for general use; and (3) statistics for special research.

A small agency engaged in some kind of social work keeps certain records of its activities. It has no intention of making any systematic statistical analysis, and the volume of data accumulated probably does not warrant such a plan. Its aim is to do social bookkeeping for its own purposes, and little formal organization is required. The administrative head or a staff worker may record information on regular forms or may keep the data in a sort of day-book. At the end of the year or at some other interval when a summary report is needed, the same individual may add up the cases, classify them according to certain attributes, and present them in written form. A larger agency or institution, like a state hospital or a prison or a family relief society in a large city, may have no intention of publishing statistics for general use or engaging in extensive statistical analysis. In other words, it may merely do social bookkeeping on a larger scale than the small agency. Where 1,000 or more individuals or families are dealt with in a year, some statistical organization is necessary, if the bookkeeping is to be of any use in showing the volume of work or in influencing policies. The day-book plan of recording is useless for a quantitative analysis of the work of the agency or institution, though as a matter of fact it is employed by public poor relief officials quite generally because they are concerned with individual cases and not with quantitative summaries of the work done. If any constructive use is to be made of the bookkeeping, regular forms for recording data must be devised and carefully filled out. The items to be recorded need to be defined carefully so that one record will be comparable with another. One or more clerks whose principal work is to keep the records and file them systematically are needed, and they should know enough, or be taught enough on the job, to enable them to tabulate the data in summary form. The data are then ready for further analysis and study, which may be done by the administrator or by a staff worker who has had statistical training. The most important step in this simple organization is preparing the record forms and defining the items

to be recorded. For several years the Russell Sage Foundation has been trying to devise a satisfactory statistical card for family case-work agencies, but so difficult is the problem that the card has been revised several times. Most demographic items can be defined rather easily, but those relating to the work of the agency or institution are often difficult of accurate definition, and they are the ones in which the agency or institution is most interested. When this step in statistical organization is satisfactorily taken, the others follow with ease.

Organizations which publish statistics for general use—public departments and bureaus and certain privately supported bureaus—must be more elaborately organized. A department of public welfare which collects statistics of the institutions and agencies in a state has to have a variety of forms. For example, if there are six state hospitals for mental diseases, it is important that all of them use the same form so that their returns will be comparable; fortunately a standard form for such institutions is widely used throughout the country, which simplifies the work of a state department. But the state department must devise forms for reports from poor asylums, township or county poor relief officials, and others for whom no standard statistical form exists. The persons who handle these forms have little or no training in record-keeping or in summary statistical reporting, and frequently they are not very intelligent; all this adds to the difficulties of the departments. Statistical clerks in the departmental offices must be better trained and able to detect and follow up errors in reports until they are corrected. Such a department ought to have a full-time statistician to supervise the collection of the material and tabulate and analyze it, but some of them do not have such a specialist. Much time would be saved and accuracy increased, if the department used punching and sorting machines, for these save labor, reduce errors in tabulation, and enable it to make much more worth-while analyses of the data collected. The Bureau of the Census at Washington has the largest statistical job of any organization in the country, since it tabulates not only the data for 122,000,000 individuals in the population census but hundreds of other series of data. It now uses 166 of these machines. The larger the organization, the more professional statisticians it requires, even though the data are not analyzed and presented as special studies.

The organization is somewhat different when the purpose is special social research. New forms must be devised for each sepa-

rate study. A basic staff of clerks and statisticians is necessary, but experts who are familiar with the special field of investigation and who have had statistical training are of vital importance. It is a common opinion among workers in the social sciences that a good research man must know his subject first, and then know statistical methods. If it is impossible to find some one possessing both these qualifications, it is better to select one who knows his field, letting him learn statistical methods as opportunity permits. However, with the increasing emphasis upon statistics in social work and the social sciences it is usually not difficult to find a specialist who also possesses statistical training. The same set-up of mechanical equipment is necessary in a research organization as in an executive department or bureau, and whatever other equipment and personnel will add to efficiency should be obtained. When funds restrict the scope of the work, as they usually do, it is better to limit the work undertaken and provide adequately whatever is necessary to do it well.

CHAPTER III

The Nature of Statistical Research

STATISTICAL method is only one of the methods by which social research is carried on. The historical method may utilize statistical data or it may not. In the past history has been largely depictive; it has given a picture of a period of social life and has shown in general the sequence of events. This has been the work of a literary artist who had a flair for discovering facts and weaving them into a coherent pattern. In more recent years, however, some historical writing has made considerable use of statistical material, notably in economic and social history. Effort has been made to write the recent history of social and economic institutions by assembling data in a sufficient quantity and of such homogeneity that the depictive method would not be of great importance. Philosophical social research usually relies upon history for its data and is merely a special form of history, or it has attempted to generalize from impressions made in the study of a wide variety of contemporary facts. Philosophical research has an important place in the social sciences as a critical tool; particularly is the criticism of the logic of the social sciences an important function of this kind of research. But the method of research which is more often contrasted with statistics is case study. Most historical writing is based upon case study, and the generalizations which have proceeded from philosophy have on their factual side their basis in assumptions about cases. Qualitative and quantitative research are frequently set out as combatants in the field of social research, but that is a mistaken conception. Professor Wesley C. Mitchell said in his presidential address to the American Economic Association in 1924, "In the thinking of competent workers, the two types of analysis will coöperate with and complement each other as peacefully in economics as they do in chemistry."[1] In this respect as much may be

[1] Mitchell, W. C., "Quantitative Analysis in Economics," *American Economic Review*, Vol. XV, No. 1, p. 12.

said for the value of qualitative and quantitative methods in social work and in the other social sciences. All valid methods of research applicable to social data will be found useful in the social sciences. It is the purpose of this chapter to set forth briefly the nature of that form of social research which is called statistical research.

I. THE NATURE OF CASE STUDY

Because in recent years there has been a good deal of controversy in the social sciences over the relative merits of case study and statistics and because many students of social statistics will also do case studies and case work, a more extended discussion of this subject is desirable. A case is an individual instance. It may be a delinquent boy, a family, a community, or even a nation. Whatever it is, it is a single unit of the kind under examination. Upon this subject Professor Giddings has illustrated a case in the following manner: "The case under consideration may be one human individual or only an episode in his life; it might conceivably be a nation or an empire, or an epoch in history. The cases with which social workers are apt to be concerned are individuals, families, neighborhoods and communities. The cases in which the ethnologists, historians, and statesmen are apt to be interested are noncivilized tribes, culture areas, historical epochs and politically organized populations. Demographers are concerned with the evolution and degeneration of populations in respect of their biological and psychological quality, and of their vitality."[2] Some of these "cases" will appear unusual to the student. But if the concept of the case as a single unit of the kind under study is kept clear, the accuracy of Professor Giddings' illustrations will be obvious. Suppose a student of population, known as a demographer, is concerned with the growth of the population of a country like the United States. The United States is a nation, and among all the nations of the earth it is a case. It is one of its kind, a single national unit with differences that distinguish it more or less sharply in population from other nations. Statistical data will be used to study this case of population growth, but the statistics are data concerning the individual human being, not nations as a class. In a statistical study of population growth of all the nations of the earth, each nation would be an item in the series, in other words,

[2] Giddings, Franklin H., *The Scientific Study of Human Society*, p. 95. University of North Carolina Press, 1924.

a case. Statistics is, therefore, a method of dealing with a large number of cases at once, or it is the method used to consider the distribution of a single trait among many cases of a similar kind.

This is perhaps the essential distinction between case study and statistical study. Statistics analyzes the distribution of cases as units or as single traits of cases. Case study analyzes the combination of all traits in a particular case. If this distinction is valid, and it is the one coming to be held by students of the social sciences, there can be no controversy over the usefulness of both methods. They have different functions as methods of research. A case study requires as complete a description as possible of all the facts conceivably pertinent to the case. Some of the facts, like age, will be objective and quantitative; others, like attitudes, will be qualitative. But both kinds are necessary to form a judgment about the case. An illustration will make this clear. It is drawn from the field of family case work. Here is a family composed of the father, mother, three small boys and two girls, the oldest of whom is twelve years of age. The man is out of work, the family is out of food, and the rent is overdue. They have asked for financial relief. Obviously relief is necessary and will be given. But it is the business of the case worker to find out why the family is in such circumstances and to devise a plan for rehabilitating it. The name, age, and sex of each member of the family are obtained. The father has no regular occupation but is a laborer and works at whatever he finds. For eleven years the family has lived in the same tenement of two rooms. Four of them sleep in one bed and three in another. The mother seems to be mentally slow, and the oldest child, a girl, appears to be mentally retarded. Is this the first time the family has had to ask for charitable relief? A check through the social service exchange indicates that it probably is, because none of the relief agencies in the city has registered it previously. One thing is clear; for a number of years the family has made its way, and that is a basis of hope for the case worker that the family may be restored to self-maintenance. But the explanation of the present condition must be found. The names of previous employers of the father are secured; later they are interviewed. One for whom he worked nine years explains that he had been a steady worker but that a few months before he was discharged he gradually became too independent and ceased to get along with his foreman. The only thing the employer could do was to let him go. The man's jobs after that had been temporary.

He had been discharged a few times and had quit several without any adequate explanation. Further investigation showed that the man had been drinking more in recent years and frequently was drunk for several days at a time. The mother and the children were examined, and the mother and two children were found to be feeble-minded. These were the facts that formed the case worker's basis for making a plan of social treatment. Some of these facts are objective and quantitative, but qualitative factors enter into the explanation of the man's behavior. Both kinds of facts are important for social case work with this family. It might even be helpful to compare this case with some other having similar characteristics in certain respects. Complete description and comparison of case with case are the two fundamental principles of case study, and they are the indispensable tools of the case worker.

The procedures of case study and of statistics have been distinguished, but it remains to evaluate case study as a scientific method in the social sciences and in social work. First, then, as to what case study cannot do: it cannot generalize. There is no valid basis in its procedure and in the facts obtained for a generalization that would apply to other cases, for the simple reason that the other cases to be generalized about have not been included in the study. Here the concept of "population," as it is used in statistics, will help to clarify the point. The term population in statistics refers not only to human population, but to any kind of objects under study. The population in a city might be the total number of people, or the total number of business establishments, or the total number of children of school age. If we are studying a problem that concerns an entire city and it is desirable to draw conclusions applicable to the whole city, either all the "population" must be included in the study or a representative sample of the whole population must be taken. The sampling method is generally resorted to, because it is reliable (see Chapter XII), requires less time, and is less expensive. If the problem concerns a particular city, then the objects for study in that city are the population. If the problem concerns a state, the objects for study are all such objects in the state. If the sampling method is used instead of a complete study of all the objects, the sample must be selected carefully. A carefully selected sample for study permits conclusions which are applicable to the entire population. But neither a study of all the objects nor a sample of this population permits conclusions about the population of another city or state.

Other cities and states were not sampled or studied as a whole. For the same reason a case does not permit conclusions which can be applied to other cases, because the single case was the population. The other cases were not studied as a whole, and one case is not enough to be called a sample. Conclusions valid for the particular case may be drawn, but no generalization can be made concerning other cases. Generalization requires the study of a large number of cases which are representative of the whole population. In view of the fact that case study involves qualitative factors which cannot be reduced to statistical data, the only way a large number of cases can be treated statistically as wholes is to regard the judgments or conclusions about cases as statistical data. This may be done, but it can be done by none but experts, and even then the results may be questioned. Hence for practical purposes we may say that case study does not permit generalization. That is a function of statistical study and is limited to generalization from quantitative data.

But case study may be an important aid to statistics, since every statistical study is preceded by case study or accompanied by assumptions about cases. Such an apparently simple matter as the enumeration of the population of the United States assumes a knowledge of cases, namely, the human beings who constitute the population. They have age, sex, occupation, place of residence, etc., which may be enumerated or measured. This is, of course, superficial case study, but the knowledge which the Bureau of the Census has of these human cases enables the officials to decide that certain items, or factors, should be counted in the census. That is case knowledge, which comes from a consideration of cases and is an aid to factorizing for statistical purposes; the factors of importance must be determined and defined. As such, case study even when superficial is an important step in all social research and particularly in statistical studies.

It should not be concluded, however, that case study is merely the handmaiden of statistics. It has a function secondary to no other method in the social sciences. By means of case study control over the development of events in individual cases is achieved for practical ends of amelioration. The aim of all scientific investigation is to secure knowledge upon which control may be based either for practical or for scientific ends. In this sense case study is highly important in social work. In so far as the dependency in the case cited above is due to social factors, the study of this case

may be expected to lead to control; if biological conditions, such as feeble-mindedness, are paramount to all others, then control of the conditions leading to dependency in this case may not be achieved, though through segregation or sterilization control could be achieved of the biological factors in so far as the next generation is concerned. Taken in this sense, case study has its own independent value as a scientific method.[3]

2. QUANTITATIVE DATA

Statistics is concerned with quantitative data: the quantitative nature of social data may arise in two ways. One way is by counting, and the other is by measuring.[4] The fundamental difference between counting and measurement is discussed below in connection with the treatment of continuous and discontinuous variables. There may be some dispute as to which social facts can be counted or measured, and which cannot be treated in this way. Some qualitative facts, such as attitudes, do not lend themselves easily to counting or to measurement, because their definition is not clear enough, though some students have tried to measure them by means of a scale. It might be questioned whether or not the number of insane persons can be counted. Certainly the number in the total population never has been counted, but we do commonly count those in institutions. Until it is possible to connect insanity definitely with neurological changes, it will be a qualitative fact, but we shall probably continue to count the insane because of the tremendous importance of the problem. The concept of juvenile delinquency is also rather vague, even in the statutes, but this difficulty is somewhat overcome by classifying as a delinquent every child brought into the juvenile court for misbehavior. Age, weight, height, population per square mile, income, children enrolled in school, criminals in institutions, families receiving charitable relief in a city, occupation, marriages, divorces, and many other social facts can be counted or measured and are generally accepted as quantitative facts. Statistical study of crime, delinquency, and insanity is of necessity restricted to the study of those individuals

[3] For a more extended discussion of the validity of case study as a method in social science, see "The Relative Value of Case Study and Statistics," by R. Clyde White, *The Family,* January, 1930, pp. 259-265; also Lundberg, G. A., *Social Research,* Chap. VIII. New York: Longmans, 1929. Also Chapin, F. S., "The Problem of Controls in Experimental Sociology," *Journal of Educational Sociology,* 1931.

[4] In this connection see Yule, G. U., *An Introduction to the Theory of Statistics,* p. 7. London: Charles Griffin & Co., Ltd., 1924.

who are legally defined as criminal, delinquent, or insane and are confined in an institution or otherwise taken under special supervision. Hence we are dealing with court cases rather than with crime, delinquency, or insanity in general. Many social facts which are essentially qualitative and not amenable to quantitative treatment are given a quasi-objective status by the fiat of a court, custom or other authority, and statistical studies are made of them. There can be no objection to this practice, but at all times it should be clear what facts are being considered. That is, the "population" must be clearly defined in the mind of the worker and in the mind of the public which may read a report of his work.

The kind of quantitative data obtained by counting is known as an enumeration of attributes. We *attribute* some characteristic to the individuals (individual does not refer necessarily to a human being but to any unit of a class of objects to be studied) being studied. The redness of an apple is an attribute, and so would be the blue eyes or light-colored hair of a Nordic. The blackness of the Negro and the yellowness of the Chinese are attributes. Nationality is an attribute: English, French, German, Russian. Because these people live in a certain political unit and speak a certain language, the name of this political unit and the language spoken is attributed to the individuals, and English, French, Germans, and Russians are counted by the census. Of course, being English or Russian implies many other qualitative facts besides nativity and language. The census is to a considerable extent an enumeration of attributes. The chief problem here is the discovery of precise definitions of the attributes. Who, for example, is a Negro? Is an octoroon a Negro? He has seven-eighths white blood and one-eighth Negro blood. Some Italians and other south Europeans have more pigment in their skins than octoroons, but they are not classified as Negroes. By convention anyone who knows he has a Negro ancestor is a Negro. On that basis the census is taken. Of course, some people who have a Negro ancestor in the distant past do not always return themselves in the census as Negroes, because in everyday life they pass as white people. Again, who is an employed person? If he is laid off for a week at the time the census is taken, is he employed? If he is on vacation, is he employed? If he works at an illegal occupation, like bootlegging, is he occupied? Once more, who is married? Is a man married, if he regularly lives with a woman but never bought a marriage license or had a marriage ceremony? Such questions indicate the

necessity of precise definition so that the attribute will everywhere be recognized and, when counted, placed in the same classification.

Another characteristic of statistical variables is that some of them are continuous, and some are discrete or discontinuous, variables. In theoretical problems this distinction should be made, while in practical work it is not so important, though it ought to be understood. "A discrete variable," says Rietz, "is one whose values differ by assigned steps, often by unity; for example, the number of children in a family, the number of kernels on an ear of corn. A continuous variable is one whose values may differ by amounts which are infinitely small; for example, the weight of a man, the temperature at a place."[5] The number of children in a family would always be a whole number; there never would be 3½ children. The assigned step, as Rietz calls it, is one child. Upon first thought interest rates might be regarded as a continuous variable, but they are not so in practice. They are fixed by custom in terms of units, halves, quarters, and eighths of per cent. An interest rate of 4.247 per cent would never occur. But the weight or the age of a man may vary in amounts infinitely small. Weight might be expressed in pounds or kilograms, or it might be expressed in grains or milligrams and fractions of these small unit measures. The continuous variable changes by amounts as small as the investigator may wish to use, and a curve representing such data could be smoothed with theoretical accuracy. The concept of discrete and continuous variables will appear again when frequency curves are discussed in Chapter VII and when probability is discussed in Chapter XII.[6]

Quantitative data which may be measured are called variables by Yule. This concept of the variable is commonly held by statisticians. Yet in usage there is an exception, when attributes are treated as variables. In line graphs (see Chapter VII) the data plotted on both the horizontal and vertical scales are spoken of as

[5] Rietz, H. L., in *Handbook of Mathematical Statistics*, edited by H. L. Rietz, p. 20. Cambridge: Houghton Mifflin Co., 1924.

[6] The complexity of the problem of variables is too great to enter into a long discussion of it here. It is perhaps the fundamental problem in higher mathematics. For a general discussion of the subject the student is referred to Russell, Bertrand, *Principles of Mathematics*, Vol. I, especially Chaps. I, VIII. Cambridge University Press, 1903. See also McMillen, A. W., *Measurement in Social Work*, University of Chicago Press, 1930; and Chapin, F. S., "The Meaning of Measurement in Sociology," *Proceedings of the American Sociological Society*, 1930, pp. 83-94, for specific applications of the subject to social statistics.

variables. Attributes are shown in frequency distributions and are plotted just as frequency distributions of variables in Yule's sense are plotted. But with this exception, which has practical justification in technical procedure, the term variable will be used to refer to quantitative data which can be measured. Such facts as age, time, price, income, physical production, and perhaps levels of intelligence are true variables and are expressed in terms of magnitude. There are others, of course, but these will suffice for illustrative purposes. Time is measured in seconds, minutes, hours, days, weeks, months, years and centuries. These are conventional divisions, but each bears a definite relation to the other; they are measures of time for purposes of convenience and suffice for practical purposes. Price and income are expressed in dollars, pounds, marks or francs, and each of these national monetary units is definite—the changing purchasing power of money is another problem. Physical production is measured in pounds, tons, ton-miles, etc. Levels of intelligence are expressed in terms of the ratio of the chronological age to the mental age. The question may be raised whether the devices for measuring mental age are actual measuring sticks, but, if it is assumed that they are, then the intelligence quotient is a true variable.

Variables may be classified as independent and dependent. This means that one series of facts to which another series is related is treated as cause, and that the second series changes in accordance with the first. Speaking of plotting variables on a graph, Brinton says, "One of the variables is used as a standard or measure by which to interpret the facts under consideration, and it may be called the 'independent variable.' The other variable, which is interpreted from the independent variable, is called the 'dependent variable.'" And further, "It is difficult to make a general rule for determining in any case which is the independent variable and which is the dependent variable. The decision depends entirely on how any set of data is approached and on the habits of mind of the investigator."[7] Considering the way in which the term variable is used in mathematics, Brinton's statement is perhaps both too broad and too indefinite. If two variables are functionally related, the one regarded as the function of the other is obviously the dependent variable. Unemployment is unquestionably related to poverty in the case of workmen. It would be possible to measure

[7] Brinton, W. C., *Graphic Methods for Presenting Facts*, p. 84, McGraw-Hill, New York, 1914.

the time each man in a group had been unemployed and to set up some scale for measuring his degree of poverty. If other factors entering into poverty are held constant, we may measure the interdependence of unemployment and poverty. The amount of unemployment in each case is the independent variable, and the degree of poverty in each case is the dependent variable. Poverty is in this sense caused by unemployment. In statistical language Y, poverty, would be a function of X, unemployment, and they would be used in that way, if the method of correlation is employed to measure the degree of interrelationship. Sometimes in plotting two series of data it may be desirable to use as Y one series which in other connections would be treated as X. But this is a practical problem and does not alter the fact that the independent variable is always in a real sense independent and that the dependent variable is always in a real sense the function of the independent variable. One of the ultimate aims of social statistics is to predict events—a goal yet far in the future. When prediction is the objective, it is the behavior of the dependent variable that we want to anticipate. The independent variable is regarded as the cause of changes in the dependent variable.

3. MULTIPLICITY OF FACTORS

In the discussion of variables the impression may have been given that the scientific study of society is simpler than it really is. The aim of pure social science is to find causes. Applied social science uses this knowledge of causes for the purpose of controlling events in the interest of human welfare. Social statistics comes to this problem in a different manner from that employed by the older social scientists. In the earlier literature of the social sciences cause was something fixed and could be discovered like the law of gravitation, but careful case studies and statistical analysis have led to the discovery of a more perplexing situation and to a humbler conception of social causation. It should be said also that case studies have contributed to this viewpoint. Why do we have a business depression? Some decades ago it would have been blamed on the government or on the bankers, but the study of business cycles has revealed the complex milieu in which a depression occurs. Social scientists hold a point of view which is not even yet held by the general public. In the middle years of the last decade it was announced through the press and from the platform that we should never have another depression. Yet one of the

worst depressions in the experience of the American nation occurred in 1929-—. The social scientists had pointed out the beginnings of this depression more than a year before it was generally admitted, but even in 1932 they could not agree on an explanation. The problem is too complex; the factors entering into it are many. There was overproduction in certain basic lines; underconsumption was cited as a cause; other suggestions have been scarcity of gold in some countries and oversupply in others, mass production, Russian dumping, decay of capitalism, a tariff that is too high, a tariff that is too low, etc. Not all of these have been advanced by sober economists, but they have been advanced by men in responsible positions. The one thing clear to social scientists is that nobody has a sufficient understanding of all the factors involved to explain the cause of this depression.

A social condition is influenced by a multiplicity of things: social, psychological, physical, and biological. Amid such possibilities about the best the social statistician can do is to record facts which seem to him important and then observe the quantitative changes which occur in these facts. The rate of change increases, and the number of cultural factors grows. That is not as simple as it sounds, because sometimes there is doubt as to what are the facts, and particularly what are the significant facts. That, of course, is a problem common to the natural sciences also, but apparently it applies with greater force to the social sciences. We count the criminals in our institutions. They are studied by isolating various objective factors like age, sex, occupation, education, place of residence, and previous criminal record. These facts are relatively easy to obtain. But are they the significant facts? If they are, what is the significant relationship existing among them? If not, are the significant facts psychological, psychiatric, or otherwise intangible and so non-statistical? Such questions suggest the meagerness of the present scientific achievement of the social sciences, but they also suggest the importance of more strenuous effort to factorize the social situation in a significant way and in a way that permits statistical analysis.

Experimentation has a limited use in the social sciences. Human beings object to being the objects of a mass experiment; they do not submit to it like electrons or guinea pigs. Yet sometimes a condition is set up which might be a social experiment for the purposes of the statistician. A plan for building playgrounds in a city might be followed by the observation of the effects of these play-

grounds on the rate of juvenile delinquency in their neighborhoods. That would not be set up as an experiment in the sense that a chemist enters his laboratory and sets up an experiment, but it would provide a new factor in the social environment, and the observer might record changes in behavior, like delinquency, which followed the introduction of the new factor. The conditions of the experiment would have been provided by the city government and for the purpose of meeting a public demand for recreation, and the social scientist would accept the set-up and utilize it for his own purposes without in any way disturbing those plans. The Eighteenth Amendment is sometimes called an experiment, even a "noble experiment," but it is a social experiment on such a gigantic scale and involves such a complex of social factors that the efforts at statistical appraisal of it have so far been questionable. Nevertheless, in spite of the fact that the social scientist cannot set up many social experiments by himself, he may make use of such experiments as the two mentioned above to try his technique and to refine his methods. A few successful studies of such experiments might lead to much more extensive efforts of city, state, and federal governments deliberately to analyze the effects of an administrative policy or of new legislation. Sometimes, of course, they do not want to know the effects, but in other cases the political values might not be so potent.[8]

The complexity of social situations serves to emphasize the need for statistical analysis. The politician, the statesman, and frequently the historian have explained social situations by "fortuitous occurrences of special or isolated character which do not appear to operate or recur in any fixed order."[9] The multifarious occurrence of the same factor in different magnitudes requires a method of summary statement provided only in the graphs, the averages, the measures of variations, and the correlations of the statistician. The factors which show a long-time development in one direction, accompanied by short oscillations and by seasonal fluctuations, are the fundamental social data; the fortuitous events should also be considered, but only in proportion to their importance in the social order. The student of social statistics needs to be fully aware of the innumerable social factors and their com-

[8] See Chapin, F. S., "The Problem of Controls in Experimental Sociology," *Journal of Educ. Soc.*, Vol. 4, No. 9, pp. 541-551.

[9] Rice, Stuart A., "The Historico-Statistical Approach to Social Studies," in *Statistics in Social Studies* edited by Stuart A. Rice, Chap. I. Philadelphia: University of Pennsylvania Press, 1930.

binations, but his attention should be directed to the discrimination of the more significant factors and to experimentation with these by the methods of statistical analysis.

4. HOMOGENEITY OF DATA

Homogeneity in statistics refers to data of the same kind. Apples and potatoes cannot be added; nor will much be known about apples by either seller or buyer if big apples and little apples, good apples and rotten apples, cooking apples and eating apples are all put into a single class. Knowledge of apples is gained by separating them according to certain attributes and then putting together those with like attributes. This homely illustration will suggest that social situations are analogous. Social facts of like kind must be put together. The greater the degree of likeness, or homogeneity, the more reliable will be the results of statistical analysis. Among sociologists the students of rural life have gone furthest in understanding their problems, and no small part of the explanation of this lies in the fact that they have factorized their problems with a view to statistical analysis and have selected their data with a careful eye to relative homogeneity. A farmer has been recognized as a human being, but he also has children who go to school, he is a churchman in many instances, he belongs to lodges and farmers' organizations, he is a citizen and participates in the government of his community, he cultivates the land and raises livestock, and he buys and sells in an open market. But he does all these things in varying degree: some farmers make the most of the schools for their children; others do not. Some have a high general standard of living, while others grade down to a very low standard. The rural sociologists have been occupied with these problems and have been able to understand them, because they have analyzed their data into homogeneous classes.

Homogeneity is a matter of degree, and the highest degree is to be desired. Any degree of heterogeneity introduces extraneous factors which have to be considered or neglected. If they are neglected, the validity of the results of the study is questionable to the extent of the influence of the extraneous factors. For example, in a study of crime in which the main interest is crimes against property, the conclusions would be vitiated if all crimes were included. Only crimes against property would be pertinent. But crimes against property are of many kinds: theft, robbery, burglary, embezzlement, arson, wanton destruction, etc. Among pro-

fessional criminals it is known that a man usually specializes in a particular form of crime against property, because in that way he can become more proficient. So it would improve the homogeneity of the data if criminals were separated according to the kinds of crimes against property which they committed. Age is sometimes important. For example, hold-up men are usually rather young, while embezzlers are more likely to be considerably older. Why is such an age division found? That is one of the problems to be studied. Further analysis leads to a greater refinement of homogeneity.

The United States Bureau of the Census intends to get enough facts about each person so that demographic studies of the United States may have a maximum of dependability. The original purpose of the census was to determine who should vote and how to apportion representatives to Congress, but anyone who answered the questions of the census enumerator in 1930 knows that many more facts are now asked. Additional facts are wanted now by the Bureau of the Census in order that we may know more about the population, but, boiled down to the lowest terms, they are wanted for various statistical purposes requiring homogeneous groups of facts about the population. The variety and uses of the census material were pointed out in greater detail in Chapter II.

The degree of homogeneity possible varies according to whether the item is a true variable or an attribute. A true variable can be measured in terms of length, area, volume, weight, time, money units, pressure, etc. Approximate accuracy is possible here, but ultimate homogeneity is doubtful. A good illustration of the difficulty is found in the practice of astronomers who take many observations of the position of a celestial body and then get the average of their observations. Greater accuracy is achieved in astronomy perhaps than in any other science, and yet repeated measurements of the same thing vary slightly. Nevertheless, it may be said that the greatest homogeneity is achieved in the measurements of true variables. Some attributes, like the number of children in families, can be presented with a high degree of homogeneity. But others offer greater difficulties: nationality, race, delinquency, mental disorder, and others. One of the reasons why social statistics has lagged behind economic statistics in the systematic collection of data is the difficulty of securing comparable, that is, homogeneous, data. Why are juvenile court statistics, or any court statistics, almost uniformly unreliable? Mainly because

of the inability to reach agreement as to what are significant attributes and what are the precise definitions of these attributes. Perhaps too much has been attempted. The desire has been to understand the intangible socio-psychological causes back of crime and delinquency. But these have not yet been so defined that two different persons will report comparable facts. Only tireless trial and error, careful observation, and accurate recording will improve the quality of such social statistics.

For the most part the degree of homogeneity attained in any collection of data is a matter of judgment. A precise definition of the items sought, in the mind of the observer, is still the best means of securing comparable data. On this point Professor Giddings says, "Obviously any fact of sort or of size, of quality or of quantity, is truly representative and therefore may without error be taken as a sample of a pluralistic field, if the difference between any other item whatsoever of the aggregate and any other item of it is negligible for the purpose in hand."[10] If the pluralistic field, that is, any number of attributes or variables of the same general kind, is relatively homogeneous, then even a small sample may be representative of the whole. In a highly heterogeneous collection of data probably no sample would represent the aggregate. Good judgment is required in selecting an accurately defined pluralistic field and in choosing the sample. Statistical analysis may help in deciding whether a group of data are homogeneous or not. If data are put into frequency classes and plotted, assuming that a sufficiently large sample has been collected from the pluralistic field, lack of homogeneity will be revealed by two or more humps in the curve. That is, the frequency distribution will be bi-modal or multi-modal (methods of arriving at the mode are shown in Chapter VIII). There are other explanations of multi-modality, but this is one to look for.

5. LOGIC AND STATISTICS

"You can prove anything with statistics," says the man in the street. Liars have been classified, and with some justification, as "liars, damned liars, and statisticians." The statistician might reply, has not almost anything been proved by the use of history or by popular myths or by theology? When such questions are raised about the logical validity of a principle or method, a momentary impasse is reached. By way of explaining the popular skepticism of

[10] *Op. cit.*, p. 83.

statistics, it should be emphasized that statistics is a method of analysis of data which may be applied to any data by anybody. The analyst may have "an axe to grind." He may be excessively desirous of proving something that has utility for him or enhances his personal prestige. Statistics as a method is impersonal, as impersonal as mathematics upon which it depends. But it may be used by anyone as a means of pleading his special case. Some questions are so stated that they cannot be answered by statistical analysis, and others involve conflicting viewpoints accompanied by different definitions of terms. Professor Wolfe has cited a number of such questions in his discussion of statistics as a scientific method: Can the railroads continue to pay high wages and at the same time reduce transportation rates? What proportion of industrial workers are getting a "living wage"? What is "normalcy"? What is "prosperity"? What is overpopulation? What legislation is socialistic? What is confiscatory taxation?

Commenting upon the type of question mentioned above, Professor Wolfe says: "It may be said, with truth, that these are questions involving standards of equity of which no objective definition can be made. Yet they are the type of question upon which legislatures and the courts and the general public are constantly passing judgment, and toward the solution of which the scientific student of social matters should be expected to contribute objective data, if not formulated conclusions. It may be said that the scientific investigator should avoid problems involving such difficulties. But the patent fact remains that if the scientist does not grapple with them the non-scientist will, with results that can scarcely be expected to be as well founded as those at which the scientist will arrive. If we cannot be objective, we must be as objective as we can."[11] It is in attempting to answer such questions as objectively as possible that the reputable statistician is sometimes charged with being able to prove just anything. Likewise it is in dealing with such questions that the tyro in statistics does prove just anything, and in the estimation of the public implicates the competent statistician. Hence the logic of statistics requires a careful definition of what statistics can do and what it cannot do. If some of these matters of equity must be tackled by the statistician, then his wisdom will be judged by his careful discrimination of objective data which he may analyze from qualitative matters not amenable to

[11] Wolfe, A. B., *Conservatism, Radicalism, and Scientific Method*, p. 247. New York: The Macmillan Co., 1923.

statistical treatment and concerning which he can have no opinion *qua* statistician. Thus, the charge that anything can be proved by statistics should be altered to read, "Persons who use statistics can prove anything." The criticism is properly of the man and not of the method.

Logic is commonly divided into deductive and inductive. Deductive logic reasons from a general to a particular proposition; it applies to a particular case a truth that is known in general. Inductive logic reasons from particular cases to a general conclusion. Both methods are used continually in the social sciences, and both must be used. The formulas of statistics are deductively derived from mathematics. They are assumed to be true. These methods are then used as tools for the analysis of aggregates of particular cases, namely, statistical data from which may be inferred a general conclusion. If the social studies are to grow and to develop toward a more scientific status, workers in these fields must use the conclusions of other investigators. Whenever such use is made of conclusions previously reached, deductive logic has been employed, the conclusions being used as a part of the data of an inductive study. Some discussions of these two aspects of logic seem to imply that they are antagonistic: that deduction is an outworn method of the Medieval Schoolmen, and induction is the bouncing boy of modern scientific methods. Such a position is untenable; they are complementary methods constantly used in scientific work. The difference between the older logic and the newer is that in recent times no general conclusions have been regarded as absolute; all are subject to modification in the light of new facts. Science criticizes the premises of reasoning as well as its conclusions. That is the mark of its special superiority over the Aristotelian logic.

The collection of data bearing on a statistical problem is a step in the inductive process. Resemblances, differences, and relationships are noted. The data are classified, and then averages, dispersions, trends, and correlations are computed. The reliability of the results turns upon the competency of the worker, and the whole process is moving from particulars to generals, guided and perhaps illuminated by much that is already known. Almost any project will involve the formulation of a hypothesis for a working base. Careful analysis of the data may demonstrate the truth of the hypothesis, or it may require its modification or abandonment. Statistical work undertaken without some kind of hypothesis is

likely to be pointless, but the hypothesis should be held tentatively, only tentatively. The prestige of the worker is not bound up with selecting correct hypotheses so much as it is with painstaking, intelligent work. Defense of an hypothesis in any degree because it is the child of the worker vitiates confidence in his work. An inductive conclusion must be inevitable in the light of the facts. Of course, it may be partly demonstrated and held tentatively for further investigation.

Statistics is liable to fall into the same logical fallacies as any other kind of reasoning or scientific work. A few of the more common ones will be indicated. Of fallacies characteristic of deductive reasoning, perhaps those known as *non sequitur* and *petitio principii* are the most common in statistical work. The phrase, *non sequitur* (which may be translated, "it does not follow"), is applied to any loose argument in which the conclusion does not seem to follow from the premises. This kind of fallacy in typical logical form is more likely to occur in connection with the interpretation of the applicability of a statistical formula to a given problem than it is in connection with reasoning about the data. An illustration stated in syllogistic form will indicate this danger:

The coefficient of correlation is a measure of the degree of interdependence of two variables.
This expression is a coefficient of correlation.
Therefore, it measures the interdependence between two variables.

So far as formal logic is concerned this syllogism is stated correctly and ought to be water-tight, but the fact is that a coefficient of correlation might indicate mere coexistence instead of interdependence. The beginner in statistics is not unlikely, however, to be overenthusiastic about the discovery of a method of showing causation between two series of social facts and to assume that every coefficient of correlation indicates causation. As pointed out above, one of the services of inductive logic has been to criticize the premises upon which reasoning is based. The major premise of this syllogism requires criticism. It should be written, "Some coefficients of correlation measure the degree of interdependence between two variables." That statement of the major premise allows for mere chance correlations, of which there are many in social statistics. The syllogism, as stated, is known from experience to be not true.

The fallacy of *petitio principii* is generally interpreted by the

phrase, "begging the question." It is the assumption that the conclusion is true without proving it. A special form of it is called "reasoning in a circle," which is the attempt to prove a conclusion from a premise, when the conclusion itself is a part of the proof of the premise. Here is an illustration of reasoning in a circle: "The increase of insanity indicates a weakness in our civilization, because, if it did not, there would not be so many insane persons." In the first part of this statement, insanity is the evidence of weakness in our civilization, and in the second part the weakness of our civilization is proof of the many insane persons. The circularity of this reasoning is obvious, but it is more subtle in elaborate discussions of a problem and is harder to detect.

Some of the fallacies characteristic of inductive reasoning, such as inaccurate observation and finding what one looks for, have already been suggested, but special attention may appropriately be directed to certain fallacies of judgment and of conception. Errors of judgment may lead to the assumption of something as a cause when it is not or to the belief that, because a certain event precedes another event, the first is the cause of the second. The problem is to make sure that the occurrence of such events in sequence is not mere coincidence but interdependence. These errors of judgment are back of most magic of primitive peoples and of superstition of more highly civilized persons. But they may easily occur in statistical work. A new high tariff is passed by Congress, and within the next few months the country is in a stage of rising prosperity; therefore, say the high tariff politicians, the tariff caused prosperity. As a matter of fact, a study of the economic history of the country for the past generation indicates that the peaks of prosperity come irrespective of whether the party in power is protectionist or non-protectionist. Another instance of this fallacy is the common statement that insanity is increasing. The number of insane in institutions is increasing, but that fact does not prove that there are more insane persons in proportion to population than there were twenty-five years ago. Fallacies of the conceptual processes occur particularly in the attempt to formulate generalizations. An attempt may be made, as it has been frequently made in sociology and other social sciences, to state a general scientific law which applies to a wide range of phenomena, before enough facts have been examined or where the data are too heterogeneous to permit any such generalizations. A more common form of conceptual error among social statisticians, however, occurs in con-

nection with prediction of the trend of a series of events. Index numbers may be computed and the long-time trend worked out. If the index has to do with industrial production, it is important in many ways to estimate what the index may be a year or two in the future. This is called extrapolation. It is an important but a precarious business. New factors may enter the business situation, as they did in 1929, which invalidate all the estimates of perpetual prosperity. The limitations surrounding the statistician in the present state of knowledge in the social sciences should be a curb to overconfidence in such predictions on the basis of a computed long-time trend of business, prices, wages, crime, or what not.[12]

6. SCIENTIFIC LAW OR SCIENTIFIC METHOD?

The ultimate aim of scientific research in any field is the discovery of regularities in the data which permit of brief statement in the form of a scientific law. A law of science is a shorthand description of regularities among a certain kind of data. The recognition of a general description of regularities of phenomena accorded by scientists the status of a scientific law usually depends upon its statement in mathematical terms. The author is not aware of any scientific law, properly so called, which has not been so stated. As familiar examples, the law of falling bodies and the law of gases may be mentioned. In biology there is Mendel's law of heredity, and, coming closer to the social sciences, the law of population growth formulated by Pearl and Reed. Many others from the natural sciences could be mentioned and some from the biological sciences. But the further the data of a science are from such elemental facts as weight, gravitation, motion, and electrons, the more difficult it is to state regularities in the simplified and absolute terms of mathematics and, hence, the less likely is the science to discover laws.

The social sciences deal with such complex data that the generalizations they make about social phenomena will rarely attain the exactness required for recognition as scientific laws. Perhaps the so-called law of diminishing returns in economics is as good an example as has yet been formulated. Seager stated this law some years ago in the following way: ". . . after a certain point has been passed in the cultivation of an acre of land or the exploi-

[12] For this discussion of fallacies, the writer has drawn upon Hibben, J. G., *Logic, Deductive and Inductive*, Part I, Chap. XIX, and Part II, Chap. XVI. New York: Scribner's, 1906.

tation of a mine, increased applications of labor and capital yield less than proportionate returns in product. . . ."[13] This law can be expressed with an approximation to accuracy in mathematical terms. Of course, it assumes that various factors will remain constant, such as rainfall, fertility, and quality of seeds used in farming. Since these things do not remain constant, the practical applicability of the law is limited, but it is an approximation to the kind of statement which in the natural sciences is called scientific law. Another illustration from economics is Gresham's Law which states that, when two kinds of money are in circulation, the cheaper money drives the dearer money out of circulation. The extent to which this occurs could be stated mathematically. There are not many such close approximations to scientific law in the social sciences. Perhaps sociology, political science, and social work can boast of none.

But statistics is an application of mathematics. If it is applicable to social data, ought we not to expect to discover some regularities of social phenomena which can be stated as scientific laws? It is possible, but statistics will in general have a humbler task in the social sciences and in social work. It will be more often of administrative value than as a means of discovering laws. The aim, of course, is ultimately to discover laws. But the more immediate ideal in the social sciences is a faithful application of scientific method to the study of their data. In the words of Karl Pearson, "The man who classifies facts of any kind whatever, who sees their mutual relations and describes their sequences, is applying the scientific method and is a man of science."[14] Scientific method, in Pearson's definition, is a way of arriving at a true interpretation of facts, their relations, and their sequences, regardless of whether or not a generalization qualifying as a scientific law is ever made. The social statistician is a scientific worker in that sense. Much valuable work of this kind has already been done, and much more will be done with the increased use of statistical methods in the social sciences and in social work.

[13] Seager, Henry R., *Principles of Economics*, p. 129. New York: Henry Holt & Co., 1913.

[14] Pearson, Karl, *Grammar of Science*, 1911 edition, Part I, p. 12.

CHAPTER IV

Working Out a Statistical Problem

EVERY statistical problem presents special points for consideration, but there are a few general matters that may be discussed as steps in procedure. There can be no statistical problem for the student or worker, unless there has arisen some question the answer to which is not immediately apparent. The question may be very indefinite at first. After it has been thought out and becomes clearer, the worker begins to think of the kind of data required to answer the question. Once the required facts are decided upon, the next problem is to gather them. This is usually the most arduous step in statistical work, because accurate, comparable data may have to be gathered first-hand. That means schedules, questionnaires, report forms, and interviews. Or the data may already have been assembled in reports, in which case the field work is eliminated, but a new problem of determining the comparability of data has at once arisen. Data which the investigator gathers from the original sources are known as primary data; those which have already been assembled by other field workers and are in published reports, perhaps for wholly different purposes, are secondary data. This kind of classification does not imply that secondary data are inferior to primary data; on the contrary, they may be better than the investigator under the best of circumstances could gather for his own purposes. For example, the reports of the Bureau of the Census contain data which are secondary for any outside student, but no individual could make for his own purposes a census of the whole population which, assuming that he could pay the cost, would be as reliable as the secondary data in the census volumes, because the technique of taking a census has been developed and improved by the Bureau for over a hundred years. But for a special problem for which no data have been collected by reliable agencies, the investigator will have to make his own field investigations.

In order to indicate more concretely the steps in working out statistical problems, two examples of such studies will be described at some length. The first will illustrate the procedure in a problem for which data were gathered by the investigator, and the second will describe a well-known study based upon secondary data.

I. A PROBLEM EMPLOYING PRIMARY SOURCES[1]

Any statistical problem is usually taken from a general field of interest. Because of his connection with Indiana University in which a number of studies of crime from various viewpoints were under way, the author undertook to study a single problem, the distribution of felonies in 1930, in Indianapolis. This problem is a mere bagatelle, when the whole field of crime, even in Indianapolis, is considered. But it is important for various reasons: (1) it is fundamental to adequate police protection; (2) it is important as one means of determining the populational sources of criminals; (3) it is important to the courts to know whether a given criminal comes from a neighborhood in which many other criminals live. Other reasons for the study of this particular aspect of crime might be cited, but these suggest why the study was undertaken.

Once the problem has been roughly outlined in the mind of the investigator, the next step is to define it and determine what divisions it may have. Only objective facts can be considered, and they must be available. What are the aspects of the distribution of felonies in a city which may be studied by statistical methods? First, crimes occur at certain places—there are exceptions, such as transporting stolen property, when the crime occurs in a succession of places, but in general the charge against an alleged criminal specifies a place of the offense. Second, criminals live, even but for a day, at definite places. They are distributed over the city, and in some places their density is greater than in others. Third, the crime may be committed at the residence of the criminal, or he may go some distance from home to commit the offense. Do some types of crimes tend to be committed near the residence of the criminal and others at varying distances? Fourth, crimes are distributed by sex, certain types being more prevalent among males than females, and vice versa. Fifth, crimes are distributed by age of the criminals. Taking all crimes together, there are age groups

[1] White, R. Clyde, "The Relation of Felonies to Environmental Factors in Indianapolis," *Social Forces*, Vol. X, No. 4, pp. 498-509.

in which criminality is more prevalent than in others. Some types of crimes are usually committed by young men or women, and other types are committed by persons somewhat older. Thus, the distribution of crime in a city may be approached from five different angles, the data for each of which are fairly objective and readily obtainable through the coöperation of the court.

How were the data obtained for the study of distribution of felonies? A record form was made out and reproduced by mimeograph. Figure I gives this form:

Year 1930 Month of *January*

Case No.	Charge	Tract of Residence	Tract of Offense	Sex	Age

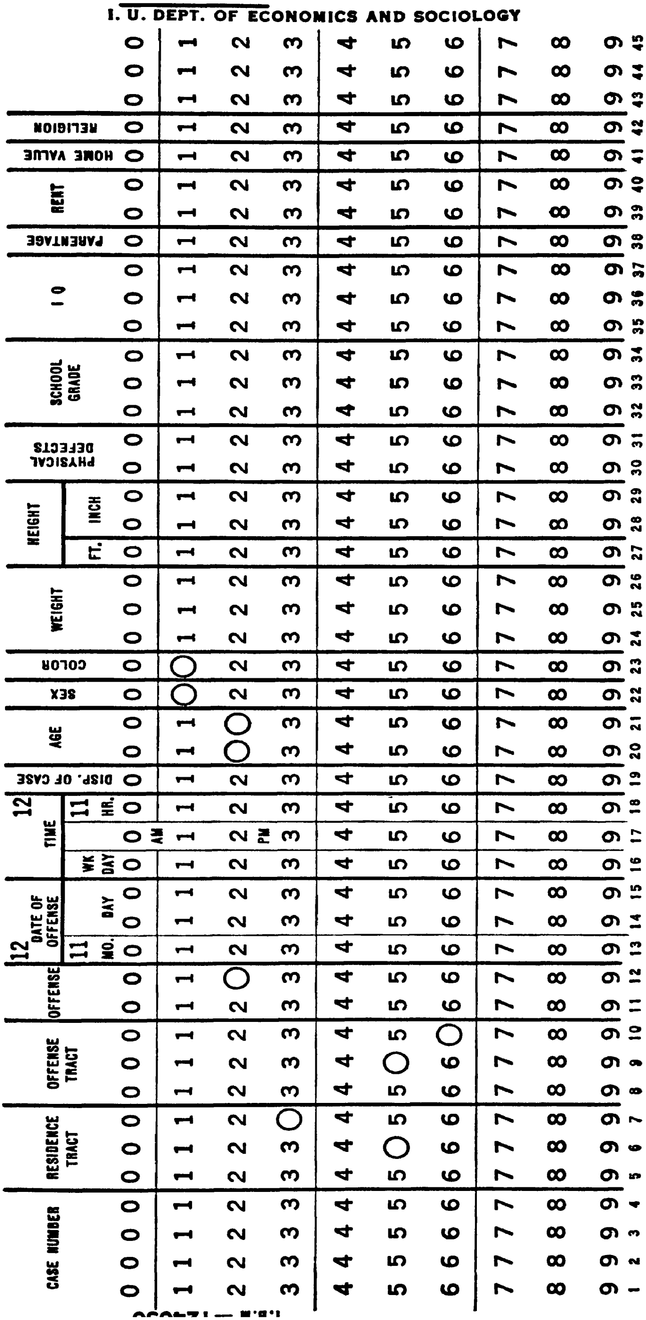
FIGURE I.—CASES DISPOSED OF BY MARION COUNTY CRIMINAL COURT FOR THE CITY OF INDIANAPOLIS

The two columns headed, "Tract of Residence" and "Tract of Offense," indicate the census tract of the city in which the criminal lived and the one in which he committed his offense. These census tracts are small areas with highly homogeneous population—homogeneous as to nationality, color, economic and social status, and age and sex distribution. Instead of taking the street number of the offender, the census tract was given. The distance of residence from the place of the offense was measured from the center of the tract of residence to the center of the tract of offense. Many offenses were committed in the tract of residence, in which case the distance traveled was negligible, because in no tract could the offender be more than a few blocks from home. Who should get this information from the criminal? It would have been a full-time job for the investigator to do that. The information is ordinarily obtained as a routine matter by the court, except for the census tract designations. The chief probation officer of the court agreed to use the forms worked out and to get the information on each

case for the study. Every case of guilt disposed of by the court during the year was included. At the end of the year the completed forms were turned over to the investigator.

After collecting the data, the next step is to assemble them in some form that permits analysis. This could have been done by hand, because the number of cases was not large, and the number of facts about each case was small. If this had been done, a work sheet would have been made up on which the age, sex, offense, place of residence, place of offense, etc., would have been tallied in, but this would have required a good deal of time. Since the investigator had access to machines for punching and sorting cards, the first step in assembling the data was to give each fact a symbol which appeared on the cards and punch the symbols out. The card is reproduced opposite.

Such cards may be worked out for any kind of statistical study, where large numbers of items and cases make hand tabulation onerous and expensive. This card was planned for a larger study of juvenile delinquency and has columns for many more facts than were obtained for the adult felons in the study of distribution of felonies in Indianapolis, but it contains columns for all the facts requested, and it could be used for tabulating the data for felonies also. The black spots indicate symbols punched out. The student will notice that at the head of each column marked off with heavy lines is printed the name of the fact to be punched in this column. Under "Residence Tract" the 3 in units column and the 5 in tens column are punched, that is, the tract is No. 53. Under "Offense Tract" the 6 in units column and the 5 in tens column are punched, which means that the offense was committed in tract No. 56. All the common offenses are numbered from 1 to 99, and it will be noticed that the 2 in units column is punched under the column headed "Offense," which means that the crime committed is denoted by the figure 2 and is assault and battery with intent to kill. Under "Age" 2 in units column and 2 in tens column are punched. that is, the offender was 22 years old. Under "Sex" 1 is punched which is the symbol for male sex, and under "Color" 1 is punched which is the symbol for white race. Each case has one of these cards, and the appropriate symbols were punched on it. The entire time required for punching the cards was about three hours. Hand tabulation would have required much longer. After a little experience of reading the symbols from the punched cards, it is quite easy to read off any data that might be wanted. However, that is



IG HO RIT CARD

not necessary, because the sorting machine which appears below does it more rapidly and with less chance of mistakes.

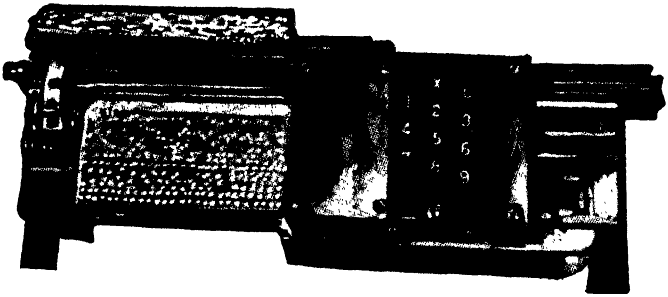


Figure III.—The Electric Key Punch

When the cards are punched, they are ready to be put into the sorting machine for any kind of classification that may be desired. Perhaps the first sorting was on sex. It is important for the analy-

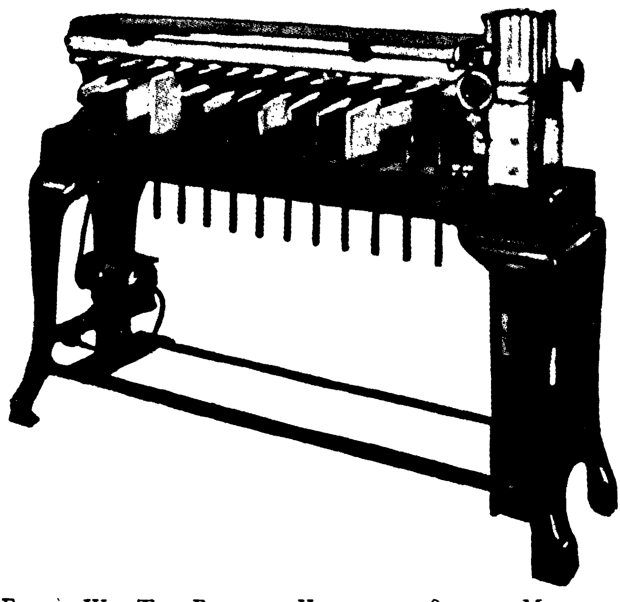


Figure IV.—The Electric Horizontal Sorting Machine

sis to separate males and females. Under "Sex" at the bottom of the column a small figure 22 is printed. This number is the guide for setting the machine to sort the sexes. The cards are put into the

feeder, the machine is set, and then the electric button is pushed. About 400 cards per minute go through the machine used—other machines sort at a higher speed. All of the cards with 1 punched, that is, males, fall into a pocket, and all the cards with 2 punched, that is, females, fall into a different pocket. Thus, male and female criminals were separated. Then the males and females were sorted by census tract, age, etc. Any kind of an analysis could be made. The cards in any particular sorting were counted by the machine and notation made of the number. When all the significant sortings had been made, the data were ready for further analysis.

Tables were then made. These tables contained the number of crimes committed in each census tract, the number of criminals living in each tract, the distribution by sex and five-year age intervals, number of crimes of each type, comparison of the characteristics of the general population with the criminals, distances from home to the place of the offense, and other cross-classifications. For the purposes of illustrating the procedure in the analysis of a statistical problem it is not necessary to present all the tables and graphs included in the completed study, but a few tables will be given to make this discussion more concrete.

TABLE I

NINE KINDS OF CRIME AGAINST PROPERTY, SHOWING THE NUMBER OF EACH, THE AVERAGE DISTANCE BETWEEN THE HOME OF THE OFFENDER AND THE PLACE OF THE OFFENSE, AND THE NUMBER OF CASES IN WHICH THE OFFENSE WAS COMMITTED IN THE SAME CENSUS TRACT AS THE RESIDENCE

Offense	Number of Offenses	Average Distance Between Residence and Place of Offense in Miles	Offenses Committed in Same Tract as Residence
Total	436	1.74	60
Banditry, automobile	9	3.43	0
Embezzlement	21	2.79	3
Robbery	20	2.14	1
Vehicle taking	76	1.77	7
Burglary	121	1.76	11
Grand larceny	117	1.53	23
Obtaining money under false pretense	38	1.47	6
Petit larceny	25	1.42	6
Receiving stolen goods	9	.90	3

There appears to be a tendency for persons committing crimes against property accompanied by violence to go farther from their places of residence for this purpose than is the case with

crimes against property without violence. The principal exception to this tendency is embezzlement. The latter is partly due to the fact that embezzlement is usually committed at a place of business; eleven of the 21 cases of embezzlement were committed in census tract 56 which is in the heart of the business district of the city. Persons who have an opportunity to embezzle funds usually have responsible positions and draw good salaries. Such persons in Indianapolis are likely to live at some distance from the business district. If this is the proper explanation here, it removes the one important exception to the inference drawn above. No case of automobile banditry occurred in the residence tract of the bandit, and only one case of robbery occurred in the same tract as the residence of the robber. If there were some way of accurately rating crimes against property according to the seriousness of the offense, a curve might be drawn to show the connection between distance traveled and the seriousness of the offense, or the degree of correlation might be computed. But this is not possible. The inference has to be made tentatively from appearances in the table.

Another interesting fact about this study of felonies is the age distribution of the criminals. This is given in the table below:

TABLE II

Age Distribution of 651 Felons Appearing before the Marion County, Indiana, Criminal Court in 1930

Age Group	Both Sexes	Male	Female
All Ages	651	631	20
16–19	194	193	1
20–24	180	173	7
25–29	74	70	4
30–34	88	85	3
35–39	49	48	1
40–44	30	30	0
45–49	21	18	3
50–54	4	3	1
55–59	5	5	0
60–64	3	3	0
65–69	2	2	0
70–74	1	1	0

The concentration of this criminal group in the ages below 25 years is striking; over half of them are less than that age. The low percentage of women is also impressive. Felonious crime is a problem to a large extent involving young men. When this age distribution is broken down to particular kinds of offenses, it is

found that crimes against property with violence and vehicle taking are committed mainly by young men. Burglars and thieves have a somewhat higher average age, and embezzlers a still higher one.

No other tables will be given, but a brief summary of other findings will be made. Crimes against the person, like assault and battery, manslaughter, murder, and rape, occur nearer the home of the offender than do crimes against property. The average distance from home of 37 crimes of this sort was .84 of a mile, and 19 of these were committed within the residence census tract. Of nine cases of manslaughter eight occurred in the census tract where the offender lived. Seven out of 16 cases of assault and battery with intent to kill were committed in the residence tract. Three out of eleven cases of rape occurred in the residence tract—this, in the matter of distance, is more like crimes against property.

The concentration of the residences of criminals is in the center of the city and especially in those census tracts where rooming houses prevail. Likewise the places where felonies are committed are in the downtown district. This is partly due to the fact that so many felonies are crimes against property, and wealth is concentrated in the downtown district. This condition is similar to that found in some other cities, notably Chicago, where studies of the distribution of crime have been made.

The essential steps in this study have been: (1) definition of the specific problem to be studied; (2) deciding upon the data to be obtained about each criminal; (3) framing a schedule on which to record the data; (4) arranging with an official of the court to obtain the data; (5) coding the data for each case so that it could be punched on the tabulation cards; (6) punching the cards; (7) sorting the cards according to various combinations of facts; (8) assembling these data on work sheets; (9) making tables of various kinds; (10) study and interpretation of the data; (11) the written report.

2. A PROBLEM EMPLOYING SECONDARY SOURCES

The general procedure in working out a problem concerning which data are drawn from secondary sources is somewhat different from that in the preceding problem. To illustrate this type of statistical problem a well-known study, *Social Aspects of the Business Cycle*, by Dorothy S. Thomas, will be used.[2] This particular study

[2] Thomas, Dorothy S., *Social Aspects of the Business Cycle*. London: George Routledge & Sons, Ltd., 1925.

has been selected for several reasons: (1) it is exclusively statistical; (2) it has been widely accepted as a good example of work in social statistics; (3) the author has depended entirely upon secondary sources; (4) the author has been under the necessity of evaluating the reliability of her material before proceeding to statistical analysis; (5) since the study was based upon data drawn from two nations and represented all of the reliable statistical data on the subject in both nations, the author has been careful to point out just what her study contributes to the knowledge of the relationships of the business cycle and other social series.

Many social scientists had noticed the apparent relationships of general economic conditions to certain social problems, but up to the time Dr. Thomas undertook her study nobody had made a thoroughgoing analysis of the problem. In order to delimit her own problem she had to examine other discussions of the subject, and in the book she has preceded her analysis by a critical discussion of previous works on social aspects of the business cycle. She says, "The subject has long been of interest to economists, sociologists, criminologists, and statisticians, but has received no wholly adequate treatment in which the relationships between these various social phenomena and the business cycle have been classified and expressed in quantitative terms."[3] Economists had noticed that marriages, births, and deaths from certain diseases seemed to be associated with the ups and downs of prices and of general business conditions. Some of them had noticed that, not only temporary dependency, but pauperism seemed to increase after a severe business depression. Others had remarked that alcoholism is a disease of prosperity. Certain crimes against property occurred more often in times of depression. Some criminologists have called attention to the fact that an increase in certain kinds of crime lagged behind the rise of prices but seemed to be connected with this economic factor; others have believed that crime fluctuates according to general fluctuations in industrial conditions rather than with the price level alone. Statisticians who have turned their attention to this problem have been concerned generally with the relation of the business cycle to marriage rates and birth rates, though now and again other phenomena have been considered.

Dr. Thomas notes increasing attention to criticism of methods of analyzing such data as time passes but concludes that none of the studies is sufficiently comprehensive to permit anything like a

[3] *Op. cit.*, p. 24.

general conclusion regarding social aspects of the business cycle. The business cycle is the term applied to the ups and downs of general economic conditions which recur every few years and which are in process continuously. The trend of economic conditions is another matter: it refers to the direction of growth over a long period of years, say, forty or fifty or more. This trend may be upward, downward, or curvilinear. Whatever its direction, it should not be confused with the short changes known as cycles or the still shorter seasonal variations. The chief criticism of method which may be made of earlier discussions of the social aspects of the business cycle is that they were really concerned with both the long-time trend and with the cyclical changes. A few later statistical studies tried to separate these two kinds of change, but no comprehensive study was made until Dr. Thomas and Professor William F. Ogburn undertook to consider all the available data in the United States and to eliminate the long-time trend from their data so that they could study only the effects of the short, cyclical changes upon social phenomena.[4] This led Dr. Thomas to undertake some more detailed study of certain American data and to supplement this with a comprehensive study of similar data in England, where social and economic statistics have been kept for a longer period than in the United States.

Since all the various social phenomena were to be compared with business cycles, the first problem Dr. Thomas had to attack was the discovery of data for, and the computation of, an index of general business. This was to be the independent variable in all cases. No single type of business could be taken as an index of general business. Several different economic series had to be combined. After an examination of various kinds of economic data, the following were selected for combination into a general index of business in England and Wales: exports of produce, Sauerbeck index numbers, percentage unemployed, production of pig iron, production of coal, railway freight traffic receipts, and provincial bank clearings. These series were selected for the following reasons: "In the first place, the series selected must move synchronously. There is frequently a difference of two or three years between the maximum of two representative series of business statistics, although both move in cycles. Series must be selected

[4] Ogburn, William F., and Thomas, Dorothy S., "The Influence of the Business Cycle on Certain Social Conditions," *Jour. Amer. Stat. Soc.*, September, 1922.

which reflect closely the general business situation, and series which are so sensitive that they forecast the general movement, as well as those which lag considerably behind, must be discarded. The series must also be as widely representative as possible of all of the most important phases of economic activity which are affected by the business cycle."[5] These conditions seemed to the author to be met by the series of business data mentioned above. Accordingly, an index was computed. It should be stated that, if such a problem as this were undertaken now, it would not be necessary for the investigator to compute an index of general business; several reliable ones have been computed and are published, currently, the most complete being those published by Standard Trade and Securities Service.

After computing the index numbers, the problem still remained to remove the effects of the long-time upward, downward, or curvilinear trend so that the cyclical changes would be uncomplicated with other types of change. Where quarterly social data were used, it was necessary to take another step and remove seasonal variations from a general business index based upon quarterly economic data. The important point for our discussion is that here is an example of rigorous effort to measure only what was intended and not a number of things that were outside of the problem as defined. Are there cycles in social phenomena which are determined in any degree by cycles in business conditions? That is the problem before the investigator. The variations left in the index numbers after removing the trend and seasonal variations are the cyclical changes. In order that these changes, whether of business, marriage rates, birth rates, crime, or other series, might be strictly comparable they were reduced to their respective units of variation (standard deviation, in this case) from their arithmetic averages. The student will not understand the full significance of seasonal variations, cycles, trends, and standard deviation until later chapters, but all that is required at this point is to recognize the use of these methods for rendering statistical data comparable in this particular problem. It is a part of the method of science—one of the rather tortuous paths to honesty in social science.

The social series of data were selected, first, because they seemed to be accurate, and, second, because it was suspected that they were affected by the business cycle. These principles of selec-

[5] *Op. cit.*, pp. 12-13.

tion at once ruled out many series, and, hence, narrowed the problem. The social series finally chosen were marriages, births, deaths, pauperism, alcoholism, crime, and emigration. Records for all these were fairly complete in England and Wales for a long period, and some of them were complete for a considerable period in the United States. Each social series has various aspects. Under marriages the relations of the business cycle to marriage rates, to prostitution, and to divorce were computed. Birth data included birth rates, illegitimacy, deaths from childbearing, and premature births. Deaths were broken down into general death rates, infant mortality, deaths from tuberculosis, and suicides. Pauperism had three phases: indoor, outdoor, and casual. Alcoholism included data on per capita consumption of spirits, prosecutions for drunkenness, and deaths from alcoholism. Crime had six divisions: all indictable crimes, crimes against property with violence, crimes against property without violence, malicious injuries to property, crimes against the person, and crimes against morals. Under emigration from England the relations of the business cycle were computed for total emigration from the United Kingdom, emigration from the United Kingdom to the United States, and the relation of British business cycles to American business cycles.

The same process of computing the long-time trend, the cyclical variations, and seasonal variations had to be repeated as for the business data. The interest of the author was in the cyclical variations of social phenomena and their relations to cyclical changes in business. Where the data were given by quarter years, seasonal indexes had to be computed and the amount of change due entirely to seasonal conditions subtracted. In all cases the average increase per year over a long period of time was computed. When the amounts of seasonal change had been subtracted, then the actual variations in crime, pauperism, and the other series from the general trend represented cyclical fluctuations. The latter are the specific data the investigator had been seeking through the long calculations up to this point. Finally, as in the case of the business series, the cyclical fluctuations were reduced to their respective units of variation (standard deviation) from their arithmetic averages. Now the business data and the social data are strictly comparable, but their exact relationships have not been calculated.

Dr. Thomas shows the relationships between the business cycle and social cycles in two ways. First, she presents the cyclical fluc-

tuations graphically, showing the business cycle and marriage rates or other social series on the same chart. Her Chart II, showing the relation of the business cycle to marriage rates and divorce rates, is reproduced below:

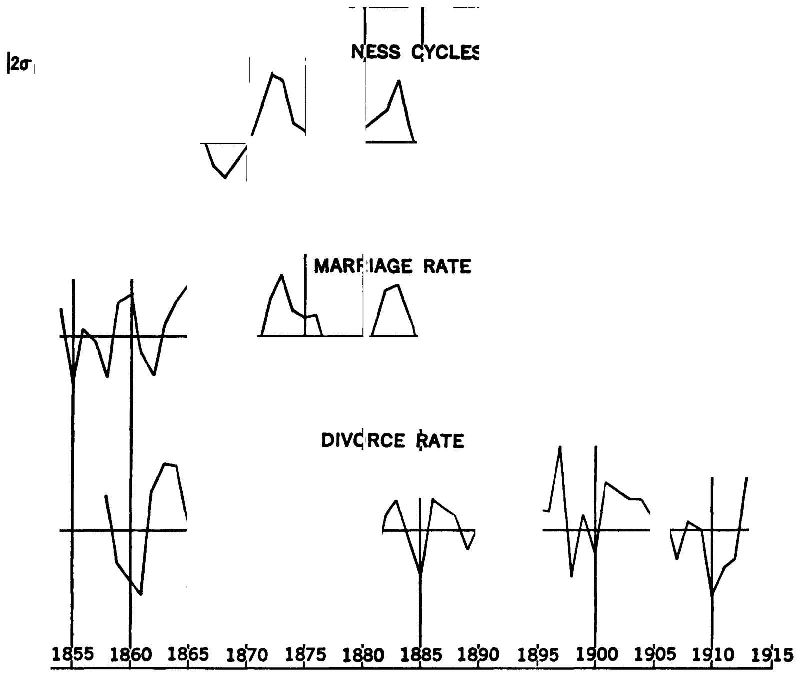


FIGURE V.—RELATION OF BUSINESS CYCLES TO MARRIAGE AND DIVORCE RATES

The ups and downs of the curve for the business cycle are closely matched by the ups and downs of the curve for marriage rates. The similarity is less marked in case of divorce, but, though divorce follows two or three years behind the business cycle, there is considerable similarity in the form of the two curves. That is, the business cycle appears to influence to a marked degree marriage and divorce rates. But the chart permits only a rough estimate of the degree of relationship. If this degree of relationship is to be measured exactly, some other method must be found. The method adapted to an exact measurement of relationships of social phenomena is correlation (see Chapter XI), and the measure of relationship is called a coefficient of correlation. This method of measuring relationships will not be discussed here in detail. It is

sufficient to state that a high degree of correlation was found between the business cycle and marriage and divorce rates—higher for marriage than for divorce. When prosperity is high, it can be expected that the marriage rate is increasing and that the divorce rate will soon start to rise, if it has not already done so. When there is a depression coming on, marriage and divorce rates can be expected to decrease. Similar graphs were constructed and similar coefficients of correlation computed between the business cycle and the other social series. The text of the book discusses the probable significance of the relationships in each case and states conclusions cautiously.

An important part of every statistical problem, after computations are finished, is its presentation in clear, concise literary form. Graphs, tables, and numerical results are included. Whether the study is to be published or not, it ought to be written up. Writing up a report of the work helps the investigator to clarify his own thinking about the problem, and it makes his work understandable to others who may be interested in it. The investigator knows more about his problem than does anyone else, and in the written presentation he can interpret his work, injecting whatever cautions are necessary. Dr. Thomas' book is an admirable example of good presentation of statistical analysis of a problem.

The steps in this statistical study of social aspects of the business cycle may be briefly summarized: (1) definition of the problem; (2) study of previous discussions of the same problem; (3) exact statement of what contribution the investigator expects to make to the understanding of the problem; (4) determination of the kind of data required to solve the problem; (5) elimination of series which appear to be inaccurate or irrelevant; (6) computation of a general index of business, followed by elimination of trend and seasonal variations so that only cyclical variations will be left; (7) elimination of trend and seasonal variations in the social series, leaving only cyclical variations; (8) comparison of cyclical variations of the business index and the various social series by means of graphs and correlation; (9) presentation of the study in literary form. Any other statistical problem requiring the use of secondary data would have its own peculiar variations from the procedure used by Dr. Thomas, but some of these steps will appear in almost any such problem.

Part Two

STATISTICAL ANALYSIS

CHAPTER V

Collection and Assembling of Data

I. DEFINITIONS

A MORE complete account of methods of collection and assembling of statistical data may now be given. In neither of the problems discussed above were all the common methods of collecting and assembling data utilized. Yet these steps are primary in social statistics. Research in the social sciences is just as strong, and just as weak, as the accuracy of the data collected. However refined and elaborate the mathematical analysis of the data may be, it is of little value if the recorded observations are inaccurate or carelessly made. A civil engineering student spends much time in the field with transit and chain, learning to make accurate observations. His measurements of elevation and distance are his data. If he fails to adjust his transit properly, errors are made. If his chain is a little short or if he does not measure exactly from his fixed points, errors render the work unreliable for engineering plans. It is no less true that the social investigator, whether he be social scientist, social worker, or social engineer, must have learned how to make accurate observations on the facts he is seeking. Furthermore, after careful consideration, he must be able to discriminate between secondary data which are reasonably accurate and those which are unreliable.

But accuracy is a relative term. It should not be thought that absolute accuracy is required in social statistics. That is an impossible attainment. In every problem there is a standard of accuracy essential to its solution. A few people are probably missed when the national census is taken, but that does not seriously impair the value of the enumeration of more than a hundred million individuals. Accuracy in the observation of attributes turns upon the degree of precision of definition of the attribute and upon the assiduity of the investigator. For example, in a statistical study

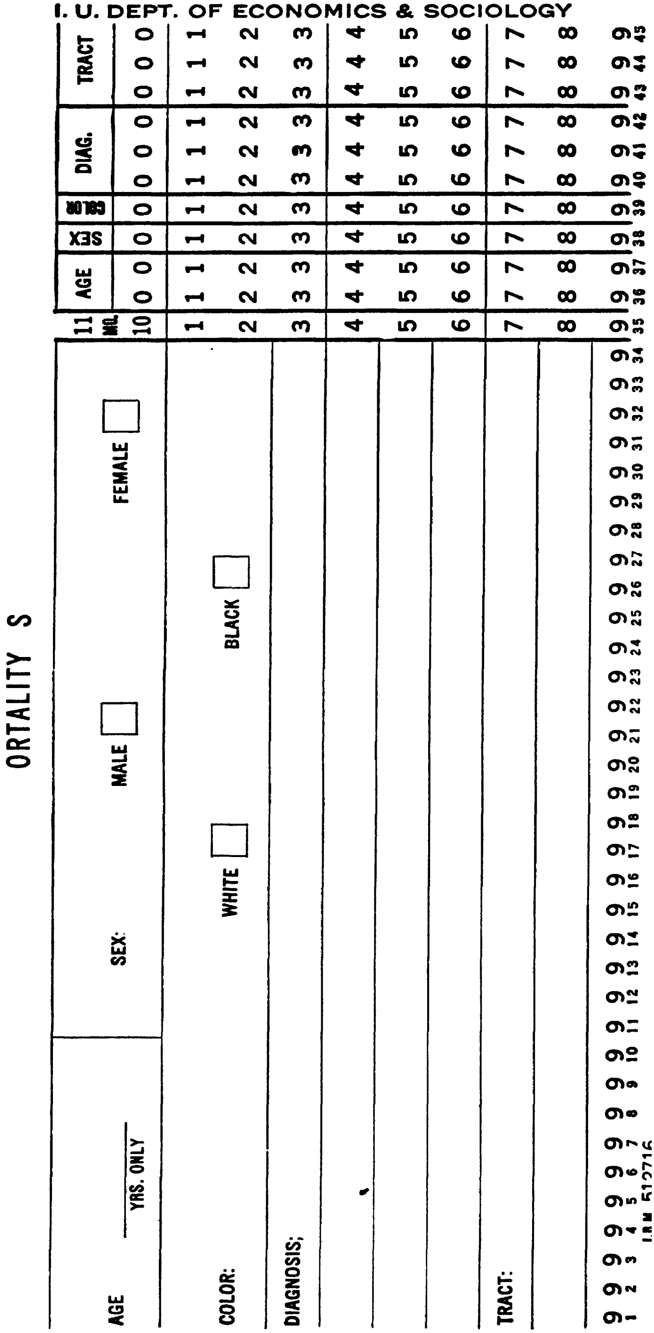
of insanity the attribute, insanity, must be carefully defined. Are only those patients who are confined in public hospitals to be considered? Or will private hospitals for mental patients provide some of the data? If they do, then types of insanity to be included will have to be decided upon, because the private hospital is likely to have a larger proportion of mild cases than the state hospital. It will not be sufficient to fall back upon the legal definition of insanity, because many people legally committable go to private institutions. Are the out-patients of mental clinics to be regarded as insane? Some of them would undoubtedly be admitted to a state hospital if application were made. From these questions, it will be clear that the standard of accuracy in a study of insanity will be arbitrary, but it will be none the less necessary in order that the applicability of the conclusions may be determined. A study concerned with true variables does not escape the necessity of a standard of accuracy. Suppose the problem is to determine the educational age of the children in a school—the educational age of a child is determined from the ratio of his school year to his chronological age. A child is eight years of age and is in the third grade. Shall we take his age in round years to the last birthday or to the nearest birthday? Or shall we express his age more exactly in years and months? If the child is being studied in December and he entered the third grade in September, should we use simply the whole number, 3, to express his grade, or should we conceive him to have moved from the round number to 3.3? This question must be decided before any data can be collected. When the standard of accuracy is decided, all the data must be collected with reference to this standard. These illustrations will serve to indicate what is meant by the standard of accuracy as a relative term.

Secondary data have already been collected. They were gathered for some purpose by the original collector. This purpose may be different from the one actuating the person now concerned with them. The investigator will collect his data in this case by assembling the publications in which the data occur. He should then determine the standard of accuracy observed in their original collection. Whatever this was, the present investigator cannot change it. If it was not sufficiently exact for his purposes, he cannot use the data. On the other hand, if he thinks the data are exact enough for his purposes, he can proceed to use them but must make only such inferences as the standard of accuracy would

warrant. He may manipulate the data in any way he wishes, but the form in which they are published will place some limitations upon him. For example, if his published data record ages in ten-year intervals, he cannot break them down and use five-year class-intervals, though he could add them and use twenty-year class-intervals. For this reason it is desirable that data, which are likely to be used by many investigators as secondary data, be published in the simplest form that anybody might want to use.

2. COLLECTION OF PRIMARY DATA

Some primary data may be collected through official agencies, provided the investigator furnishes the report forms. This is a common occurrence when a public agency is interested in a piece of research being done by an outsider. The agency will agree to order the reports made in the form desired for their research project. In such cases the terms must be the simplest possible. If there is any question about the exact definition of terms, a list of definitions must be given to those who record the information. The tabulation card reproduced below is an example of a report form, where the meaning of the terms was so clear that no list of definitions was necessary. This card has spaces on the left-hand end for the information to be written in by a clerk of the Indianapolis Department of Health which was coöperating with the author in this study of mortality. The Department of Health is asked to give the month of death, the age in years (date of birth is not always obtained on the physician's certificate, but the age is given), sex, color, diagnosis, and census tract. This particular card is an example of simple machine tabulation, because the information to be punched is written on the card itself; this can be done, if the number of items is small and sufficient space is left on the card for recording the information. The department clerk should make no mistakes in transferring the information from the physician's report to this card, because the terms are simple, objective, and capable of no misinterpretation. The only question of definition that ever arose was whether persons who lived out of the city but died in the city should be reported; these were not wanted, because this was a study of the mortality of inhabitants of the city of Indianapolis. All the deaths of residents of Indianapolis occurred in a census tract, or, if at a hospital, then the person had a residence in a census tract, and it was the residence tract that was wanted.


I. U. DEPT. OF ECONOMICS & SOCIOLOGY
ORTALITY S
AGE
YRS. ONLY
SEX:
MALE
FEMALE
COLOR:
WHITE
BLACK
DIAGNOSIS;
TRACT:
TRACT
DIAG.
COLOR
SEX
AGE
I.B.M 512716

The tabulation card reproduced on page 102 is more complicated. It could not be given to the juvenile courts for entering the information desired. A report form, embodying the same items, was printed and sent to the courts. But the terms are not as unambiguous as those in the mortality study. Every term had to be defined with precision; a few terms are rather obvious in meaning, but some explanation was necessary in all cases. "Offense" was to be called by its legal name. "Disposition of Case" was to be stated specifically: if the child was sent to an institution, the name of the institution was asked for; the case might be dismissed, or damages for property ordered paid; the child might be put on probation, or the case might be unofficial, in which case the word "unofficial" was written into the form for disposition of the case. "Age" was to be given to the nearest birthday. "Weight" was to be expressed in pounds, but "height" was to be expressed in feet and inches. Etc., etc. Even with specific definitions provided to the probation officers of the courts, some question was continually arising about special cases, or some official could not understand what was meant.

Every business or social agency, public or private, which collects information about its work is faced with the same problem. The information believed to be important and reportable must be asked for in as simple language as possible. Terms must be explained carefully and sometimes often. The collection of such data is the first element in social bookkeeping, and it is basic to statistical analysis. One of the most important functions of state or city departments of public welfare and departments of health is the collection of statistics. Some of these statistics are collected monthly, some quarterly, and some annually. The department usually has authority to prescribe the form of the report. If it has competent administrators, they want the reports to show the condition of public welfare and health work in the state. This requires certain facts which must be requested in simple, objective form. The more questions that can be answered by "yes" or "no" or by numbers, the better the report form. Of course, careful definitions of the items reported in numbers are necessary in order to secure comparable data. The report form for which the juvenile delinquency tabulation card (see Fig. II) was designed was adopted as the official monthly report form of the Indiana State Probation Department and prescribed as the form on which the courts should make their reports.

C. G. Form 19. Agent's Report to Board.

BOARD OF CHILDREN'S GUARDIANS OF________________COUNTY
REPORT OF AGENT FOR THE MONTH ENDING________________

1. CHILDREN BOARDED WITH MOTHERS:	Boys	Girls	Total
a. Placed during month	______	______	______
b. Discontinued during month	______	______	______
c. Number remaining last day of month	______	______	______
Number of mothers boarding their children			______

2. FOSTER HOMES:
   a. Applications for children:
      Number received .... ______
      Number investigated .... ______
      Number approved .... ______
   b. Children placed during month:
      By Board of Children's Guardians .... ______
      By Board of State Charities .... ______
   c. Number of children in foster homes subject to visitation .... ______
   d. Number of children in foster homes visited during current month found to be getting on well .... ______
   e. Number dropped from rolls:
      Adoption .... ______
      Death .... ______
      Marriage .... ______
      Over age .... ______
      Others (specify)

3. INSTITUTION CARE:	(Beginning of Month)	(End of Month)
Wards of the Board in the following named institutions:	______	______

4. SUMMARY OF WARDS FOR LAST DAY OF MONTH:	(End of Month)
1. In mothers' homes	______
2. In foster homes	______
3. In institutions or boarding homes	______
Total	______

5. FINANCIAL STATEMENT:	Number	Expense
a. Children boarded in own homes during month	______	$______
b. Children boarded in institutions during month	______	$______
c. Children boarded outside institutions	______	$______
Total amount contributed during month by parents for support of children		

6. AGENT'S ACTIVITIES:
   Has every ward in mothers' homes been visited during the month?
   If not, which have not been visited?

   Give reason________________

   Total number of visits to homes....
   Total number of office interviews...

7. Miscellaneous work (specify):

(Signed)_

FIGURE VII.—REPORT FORM USED BY THE BOARDS OF CHILDREN'S GUARDIANS, INDIANA

Because such a large proportion of social data are collected by official agencies and because many students who expect to take positions as statisticians will be associated with public departments, two samples of official report forms are reproduced here. The first, reproduced above, is the monthly report form used by the agent of county boards of children's guardians in Indiana, and the second is the form filled out for admission of patients to the out-patient department of the Indianapolis City Hospital.

INDIANAPOLIS CITY HOSPITAL — Form D7

Out-Patient Department

Name__________ Age __________ Race {W / N}

Address__________

Ref. by__________

Reason for referring__________

No. in family__________Adults__________Adults working__________

Children in School__________Children under School age__________

Children working__________Dependents__________Occupations of those

Employed__________

Income from Father__________Mother__________Children__________

Others__________

Expenses: Rent__________Insurance__________Installments__________

Food__________Fuel__________Others__________

Remarks: __________

Signed_

FIGURE VIII.—REGISTRATION FORM

Closely allied to the type of reports received by public agencies are the records kept by private social agencies. These agencies may not keep their records primarily for the purpose of reporting to a central collecting agency, but they keep records for their own use. A settlement house carries on a variety of activities, and the workers, as well as the board of directors, want to know periodically what participation there has been in the different activities. A public health nursing association is interested in the number of different types of cases it handles, the cost of cases, and their location in the city. Its interest in statistics may be entirely administrative; there may be no definite interest in statistical research. But statistics are indispensable to the effective administration of

the public health nursing association. Most such private agencies are related in some manner to a national organization, one of whose duties it usually is to develop standards, including standards

SERVICES RENDERED

✓ Service completed

Surgical appliance secured
Trade training for handicapped
Inst. care—aged or handicapped (excl. of a

Psychiatric e
Non-custodial psychiatric or neur. tr
Adjustment within family group

34. Mental defectiven

37. Drug habit
38. Irregular sex re
39. Personality or behavior problem (excl. of ab
40. Attitude producing conflict bet. husband &

. Non-contributing eb

1. Transportation or deport. to other loca
2. Temporary shelter
on made with neighborhood o

Printed in U. S. A.

FIGURE IX.—STATISTICAL CARD

of statistical reporting. For several years the Family Welfare Association of America has been experimenting with various statistical cards. This Association is concerned primarily with family

case work as treatment, but it is well aware of the desirability of reducing to statistical form all the data which lend themselves to such recording. Some of their data are qualitative and cannot be enumerated satisfactorily, but much of the useful information can be checked on a card. Then at the end of the year it is possible for the society to make a statistical summary of its work. A few of the larger societies are employing statisticians whose business it is to analyze the statistical data on the form card which each case worker keeps for each of her cases.

The card which the Family Welfare Association is now recommending to its member agencies is reproduced below.

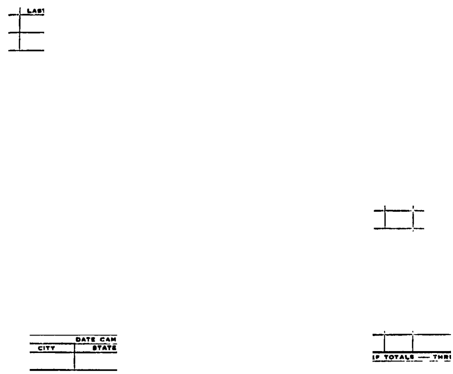


FIGURE IX-A.—REVERSE SIDE OF FIGURE IX

A questionnaire, technically defined, is a blank form mailed to the person who is expected to furnish the desired information. The response of the person interrogated is wholly voluntary. He may fill out the questionnaire and return it, or he may throw it in the wastebasket. A questionnaire may ask for information that is a matter of opinion and not capable of statistical expression. This kind of questionnaire is not under consideration here. Government bureaus and departments frequently do not have authority to compel the reporting of certain information which they want; in such cases they resort to the questionnaire method of collecting their data. This method is also used widely by private individuals and organizations having no official status. As stated above, the replies

are always voluntary; and responses are usually received from only a small percentage of the questionnaires mailed out. Government bureaus using this method probably get a higher percentage of replies than do individuals or private organizations, because the citizen is likely to feel some obligation to respond to a request of the government. The questionnaire should be so framed as to

FORM 25[1]

U. S. Department of Labor, Bureau of Labor Statistics, Washington

Dear Sir:

The Bureau of Labor Statistics is endeavoring to keep as accurate a record as possible of all strikes and lockouts in the United States as they occur. We shall, therefore, greatly appreciate your courtesy in furnishing as much as you can of the information listed below, relative to the strike or lockout here indicated.

An Addressed envelope on which no postage is required is inclosed for your reply.

Very Respectfully,

Commissioner of Labor Statistics

SCHEDULE OF INQUIRY

1. State________ 2. City or town____
3. (a) Industry________(b) Occupation____
4. Strike or lockout?________
5. Name of establishment (if more than one, give number)_
6. Date of beginning________ 7. Date of ending________
8. Number of employees involved. Male______Female______Total_
9. Cause or object, briefly stated________
10. Result, briefly stated________
11. If ordered by a labor organization, please give name_
12. If settled by arbitration, please name Board______
13. If terminated by a written agreement between employer and employees, will you kindly inclose a copy of the same? ________

[1] United States Bureau of Labor Statistics, *Methods of Procuring and Computing Statistical Information of the Bureau of Labor Statistics*, Bulletin No. 326, 1923, p. 38.

FIGURE X.—QUESTIONNAIRE OF THE U. S. BUREAU OF LABOR STATISTICS

make the person feel he has some interest in the subject investigated. This may be done by an explanatory note at the top or at the bottom, or in a letter. As few questions as possible should be asked. A short questionnaire may be filled out in a few minutes by the person who has the information. On the other hand, some questionnaires contain several pages and dozens of questions which would require several hours of work to answer conscientiously, and

few people will trouble to fill them out. If 10,000 questionnaires are mailed and only 1,000 are returned, there is always serious doubt whether the returns are sufficiently representative to be worth anything. For statistical purposes, the questions should lend themselves to "yes-or-no" answers or to answers in figures; opinions should be excluded, because they are non-statistical in nature.

Two samples of questionnaires are given to illustrate the method of asking questions.

Fig. X is put out as an official government form, but it is really a questionnaire in view of the facts that it is mailed to the person who is to furnish the information and that the response is voluntary. The purpose of it is stated in a short letter at the top of the questionnaire. The questions are few and require simple, objective answers. Questions 9 and 10 are the only ones in any way involving matters of opinion, and usually both the employer and the employees in a strike or lockout have a reason that can be stated briefly. The questionnaire is mailed to both parties to the strike or lockout. If there is disagreement as to the cause or result, further inquiry can be made.

The next questionnaire is also put out by the Bureau of Labor Statistics in connection with its current statistical record of industrial accidents. This form asks for information bearing on the amount of exposure of the employees to possibility of accident:

FORM 26[1]

REPORT OF EMPLOYMENT

Company........................Plant........................Year..........

Department	Total Hours Worked by All Men as Shown by Time Books	If Total Hours Are Not Available, Report as Below		
		Average Number Employed	Days Department Was in Operation	Usual Length of Day or Turn

[1] *Op. cit.*, p. 39.

FIGURE XI.—QUESTIONNAIRE OF THE U. S. BUREAU OF LABOR STATISTICS

Another form is used for obtaining the number of persons injured and the amount of disability. Form 26 enables the Bureau to compute the liability to accidents, and with the other data obtained on the next form in its series (Form 27) it can estimate the increase or decrease of industrial accidents over a period of time.

The survey schedule is similar to a questionnaire, but it is used differently and may be more complicated. A field worker takes the schedule and interviews the person who is to give information. The form used by the Government to take the national census is in fact a survey schedule, though it is not referred to as such. Surveys of farm houses have been made by the United States Bureau of Agricultural Economics. This Bureau is continually directing surveys in different parts of the country in connection with its studies of farm production and the standards of living of farmers. The land grant colleges carry on numerous surveys of rural communities or counties, or even surveys of some aspect of rural life on a state-wide basis. One of the most widely known urban surveys was the Pittsburgh Survey made in 1909-16. This survey was made with particular reference to the standard of living and working conditions of the industrial workers in and around the city of Pittsburgh. Recently the Russell Sage Foundation has published a directory of over two thousand social surveys which have been made in different parts of the United States.[1] In probably all of these the survey schedule has been an important means of recording the data necessary to analyze the problems under consideration. Certainly it is true of those that were carefully planned and executed. "The schedule used by the field worker," says Chapin, "is a mechanical device which is designed to provide him with a method of limiting or controlling his observation and of standardizing the method of recording that observation. In so far as inquiries on the schedule are put in a form which can be answered by a numerical or quantitative statement or by 'yes' or 'no,' the subjective characteristics of the field worker which may bias his opinion are eliminated."[2] Chapin gives detailed descriptions of field work procedures in this work. The questions must be framed carefully so that they elicit nothing but objective replies, as Chapin suggests. The fact that a field worker carries the schedule and obtains answers to the questions on the schedule

[1] Eaton, Allen, and Harrison, Shelby M., *A Bibliography of Social Surveys.* Russell Sage Foundation, 1930.

[2] Chapin, F. S., *Field Work and Social Research,* pp. 49, 50. New York: The Century Co., 1920.

by talking with persons who are familiar with the facts makes this method of securing information more reliable than the questionnaire method, and it insures responses from a much higher percentage of persons. The questions asked in a questionnaire, if at all complicated, are open to as many interpretations as there are persons replying, whereas, if the field worker has some bias which careful formulation of the schedule cannot entirely nullify, all schedules have the same bias.

A good schedule used for the study of a social condition or situation reduces considerably the necessity of having highly trained field workers. If the investigator knows what he wants and if he wants something that can be objectively defined and studied by means of objective facts alone, he can organize a staff of untrained workers to gather the material. This is regularly done every ten years by the Bureau of the Census which conducts the most comprehensive of all surveys, the enumeration of the composition and characteristics of the population of the nation. In 1930 the Committee on Compensation for Automobile Accidents, under the auspices of Columbia University, conducted a survey of persons injured in 1928 and 1929 in automobile accidents in several different states. A few trained workers were used to direct the work in each locality, but much of the calling upon families was done by college students who had had no experience in making surveys. This was possible, because the questions asked for simple matters of fact, and all the field worker had to do was to be reasonably courteous, enlist the interest of the persons injured or their relatives, and record the answers to the questions on the schedule.

Two schedules are reproduced below to illustrate the kind of questions that should be asked and the manner of asking them. The schedule used by the Committee on the Study of Compensation for Automobile Accidents calls for a great deal of information. Most of the questions could be answered with a high degree of accuracy. In practice it was found that the questions concerning the expenses of the injured person could not be given precise answers, and resort had to be made to estimates of the amounts under different headings. The reliability of the data depended upon the ability and willingness of the injured person or some member of his family to answer the questions. The field workers rarely found any difficulty in getting him to talk. The schedule suggests another thing: that is, the complexity of such an apparently simple social problem as compensation for automobile acci-

SCHEDULE FOR THE STUDY OF COMPENSATION FOR AUTOMOBILE ACCIDENTS

1. File #________________ 2. Date of accident________________

3. Name____________________ Age________ Sex_

4. Address______________________________

5. Type of accident__________________________

6. Injured was: pedestrian________ owner driver________ driver.
   or a passenger________ who was: owner____________
   member of owner's family_
   member of driver's family_
   guest of owner________
   guest of driver________

7. Driver was: owner________ owner's friend________ owner's chauffeur_
   member of owner's family____________ renter________

8. Fatal: immediately________, after________hours, ________days, ________weeks

9. Injury____________________________________
   ____

10. Date of investigation____________ Injured is M._____S._____W._____D._____

11. Occupation when hurt__________________ Earnings $____________week

12. Customary occup. during previous year____________ Earnings $__________week

13. When struck injured was: on way to or from work________________
    out for recreation____________ stealing a ride_
    other (state explicitly)________________

14. Injured was struck by: hit and run driver________ intoxicated driver________
    out of state car________________ stolen car________________

15. Disability:
    In hospital:________ emergency treatment only; ________days, ________weeks
    No disability__________ temporary__________ able to resume regular duties__________ days__________weeks after accident.
    Permanent total (state injured's condition)________________
    ____
    Permanent partial: period of temporary total disab. ________________
    Injured's permanent condition________________________
    Earnings since accident $____________________week.

FIGURE XII

16. Expenses: (If any treatment was free, please indicate)

		Paid	
Hospital..................	\$________	\$________	by________
Medical (doctor, nurse, drugs, X-ray, etc.)..............	\$________	\$________	by________
Wages of substitute for injured	\$________	\$________	by________
Lost wages................	\$________	\$________	by________
Funeral....................	\$________	\$________	by________
Property damage...........	\$________	\$________	by________
Other......................	\$________	\$________	by________
Total	\$________	\$________	

17. Compensation:

Vehicle which struck injured was insured______ not ins.______ not known__

Vehicle in which injured was riding was insured____ not ins.____ not known_

Obtained by verdict________________________________

Settlement, through efforts of attorney, with Ins. Co._

driver______________ not known_

Direct settlement with Ins. Co.________ owner________ driver_

Received from Workmen's Compensation Fund______________

Compensation received______________days/weeks after accident.

Total recovery \$______________________________

Injured received \$__________ less \$__________ paid toward expenses.

Attorney received \$__________ less \$__________ paid toward expenses.

From Work. Comp. Fund \$____________ per week for __________weeks.

18. Pending:

Sui ____

Negotiations with Ins. Co.__________ owner__________ driver__________

Recovery likely: yes______________ no______________

19. No Recovery:

Nothing sought by injured_

Claim refused by attorney_

Claim ignored by owner______________ by driver_

20. Reasons for No Recovery:

Injured was struck by a Gov't. or City vehicle________

Party II had "influence" at hearing______________

Financial irresponsibility of Party II_

No witnesses______________

Contributory negligence of Party I_

Minor injury______________

Lack of funds to initiate proceedings_

FIGURE XII.—Continued

Ignorance of recovery procedure________________
"Just did not bother"________________
Other________________

21. Received Aid From:
Life Ins. $__________ Accident Ins. $__________
Benefit Societies $__________ per week for __________weeks
Other________________

22. Family Situation:

Relationship	Age	Occup.	Salary	Contrib. to family income

House rent $__________a month. House owned clear__________
Buying house, payments $__________ a month__________
Lodgers:__________ pay $__________a week.
Boarders:__________ pay $__________a week.

23. Effects of accident:
Wife or mother went to work_
Family borrowed money $__________ from_
Used savings $__________ Other__________

24. Has injured ever been in a m.v. accident before?__________ How often?_
As driver__________ As passenger__________

25. Short story showing type of home and condition of family. (Use reverse side of this sheet if necessary)

Interviewed________________ Investigator's initials__________

FIGURE XII.—Continued

dents. Because of this complexity, the framing of the schedule required much time and discussion before it was finally adopted.

The next schedule was used for one part of a study of dependent and delinquent children in North Dakota and South Dakota by the United States Children's Bureau. This schedule was used in the study of children under the care of institutions and agencies: Notice that when an answer requires opinion, such as "interest of relatives," an objective indication of interest is suggested, namely, the frequency with which the parents visit the child. Physical, mental, and behavior characteristics are less exactly definable than are some other facts; so the field worker was to indicate whether the answer was a result of examination by a qualified person or

not. The implication is that, if not by examination, then the diagnosis is less trustworthy.

I. CHILDREN UNDER CARE OF INSTITUTIONS AND AGENCIES[1]

Schedule No. ______

Institution or agency______ Date______

Name of child______ Sex______ Race______

Date of birth______ Age______ Birthplace______

Date received______ Source______ Perm. or Temp. Care_

Nationality—Father______ Mother______

Reason for receiving______

Maintenance: by State______ County______ Family_

Agency (specify)______ Other__

Length of time in inst. or under agency care (dates, etc.)__

Disposition (specify in chronological order, placements, parole, released, adoption, boarding, with relatives)______

Interest of relatives (frequency of visits to inst.; agency visits to child's original home and present home, etc.)______

______

Family and home conditions: (check correct answer)

Mother—dead, living, married, widowed, divorced, separated, deserted

Father — " " " " " " "

Economic conditions______

Other_

Child's characteristics:

Physical______

Mental______

Behavior______

(Specify if examination)

Child's social history:

School attendance______

Dependency______

Delinquency______

Present address of child______ Removed from_

Present address of parents______

[1] U. S. Children's Bureau, "Dependent and Delinquent Children in North Dakota and South Dakota," *Publication No. 1.* 60 P. 124.

NOTE: As published in the bulletin the schedule was rather condensed. It has been expanded here to somewhat the form which the field worker might use.

FIGURE XIII.—SCHEDULE USED IN A CHILD WELFARE STUDY

In connection with forms for securing information for later statistical analysis, the score-card should be mentioned. It is a sort of schedule, but is extremely simple and not susceptible of mis-

interpretation. All the questions are answered by "yes" or "no." No answers are checked. Each item of the score-card is assigned an arbitrary weight so that a measure of the relative importance of the items may be obtained and used in the analysis of the problem.

3. ASSEMBLING DATA

After the data have been collected on questionnaires, schedules, or official forms, they must be assembled for analysis. Tabulation is not a single step but includes everything from assembling the data punched on a tabulation card or tallied on a sheet of paper to the final form of frequency or other tables. In this chapter we are concerned only with the preliminary step, namely, assembling the data punched or tallied on a work sheet (frequency distributions and setting up tables will be discussed in the next chapter). The tabulation card and the machines have already been described (see pp. 101-106). There are other machines which record classifications, totals and sub-totals by means of a printing device. The machine arranges the cards in any order desired, but, with the sorting machine alone, the operator has to count the cards either by hand or on the machine. Then the total items are recorded on a work sheet. For example, in the problem of crime discussed in a previous chapter, one of the things wanted was the number of criminals living in each census tract of the city. Each of the 108 census tracts had a symbol on the card and was punched. If the machine is set on column 7 (see p. 85), it arranges all the cards in numerical order for units place. After they are run through, the cards are gathered up from the pockets of the machine, one pocket having all cards with the figure 1 in units place, another those with figure 2, etc. They are kept in order, placed in the machine, which is set on column 6, and they are run through again. Now they are in order for both units and tens places. Once more gathering the cards up from the pockets, the machine is set on hundreds place, and they are run through again. Now the cards are arranged according to census tracts from 1 to 108. The operator may count the cards for each tract by hand, or, if a large number of cards come in one tract, they may be counted on the machine. A work sheet has been prepared with the tract numbers arranged vertically on the left-hand side, and spaces to the right are left to record the number of items in each tract. Ages may be tabulated

in the same way. The following is a work sheet for recording residences and places of offense:

Tract	Criminals Living in Specified Tract	Offenses Committed in Specified Tract
1	4	4
2	6	6
3	1	1
4	2	4
5	1	1
6	0	5
7	5	4
8	9	0
9	2	1
10	5	4

FIGURE XIV.—WORK SHEET FOR ASSEMBLING CRIME DATA SORTED ON A TABULATING MACHINE

This work sheet does not differ from some tables. If some other items were tabulated, they could be given in any detail desired, and then tables could be made up to group them in different ways.

Hand tabulation would be different and more laborious. Punch cards would not be used at all. The worker would tabulate directly from the schedule, questionnaire, or official report to the work sheet. Sometimes the data are transferred to small cards, substitutes for machine cards, for convenience in hand sorting. The following work sheet will illustrate this procedure:

Tract	Criminals Living in Specified Tracts	Offenses Committed in Specified Tracts
1	IIII	IIII
2	~~IIII~~ I	~~IIII~~ I
3	I	I
4	II	IIII
5	I	I
6	0	~~IIII~~
7	~~IIII~~	IIII
8	~~IIII~~ IIII	0
9	II	I
10	~~IIII~~	IIII

FIGURE XV.—WORK SHEET FOR ASSEMBLING CRIME DATA—HAND AND TALLY METHOD

This method of transferring the individual items from the schedule or questionnaire to a work sheet, which is the first step in tabulation, is called *tallying in.* When only a small number of

items are involved, this method is satisfactory. The larger the number of items to be tabulated, the more time-consuming and expensive it is. But machines are not always available to students or research workers, whereas this method can always be followed.

4. EXERCISES

1. Take a published piece of research, selected by the instructor or by the student, read it carefully, and list the steps in the procedure from the formulation of the project to the written report.
2. Draw up a form for an official report:
   (a) For a probation officer who has to assemble information for the juvenile court judge on a child who is to appear in court.
   (b) For the principal of a school who has to report to the superintendent attendance for the week at her school.
   (c) For a public poor relief official who has to report his cases monthly to a board of commissioners.
3. Draft a questionnaire to be sent to ministers in connection with a study of religious education.
4. Draft a schedule for a survey:
   (a) Of housing conditions in your city or community.
   (b) Of boys selling papers on the street.
   (c) Of children attending neighborhood motion picture shows.
   (d) Of delinquent girls brought before the juvenile court in a certain year.

5. REFERENCES

Chaddock, R. E., *Principles and Methods of Statistics*, Chap. XIV.
Chapin, F. S., *Field Work and Social Research*, Chaps. III, IV, VII.
Lundberg, G. A., *Social Research*, Chaps. VI, VII.
Schluter, W. C., *How to Do Research Work*.

CHAPTER VI

Tabulation of Statistical Data

I. TABULATION AND CLASSIFICATION

TABULATION has two meanings: first, the transfer of data from original schedules or reports to a work sheet or a machine card; and, second, the arrangement of data in tables. The first use of the term is due largely to the introduction of machines which were called by the manufacturers *tabulating* machines. Some of these machines do actually print summaries of the data as the cards are sorted, but the sorting machines simply arrange the punched cards in order, and the totals have to be written down by hand in some form of a table. In order to distinguish these two processes, the first one was discussed in the preceding chapter and referred to as *assembling* statistical data. The second kind of tabulation is the subject of the present chapter.

Logically tabulation is the fourth step in the study of a statistical problem for which data have been gathered by the investigator. Classification is the first step. This has to be roughly done before the schedule, questionnaire, or report form can be devised. For example, it is proposed to study the distribution of felonious crimes in Indianapolis. What classes of data are required to describe the distribution? Whatever data are needed for this purpose must be asked for in the schedule or report form. Distribution may refer to geographical distribution of all felonies without regard to type of offense, or it may refer to the distribution of the residences of the criminals only. On the other hand, it may refer to distribution of felonies by type of offense and by place of offense, or distribution may be by age, sex, race, nationality, and time also. In the study made in Indianapolis for 1930 distribution was conceived in terms of types of offense, residence of the offender, place of the offense, age, and sex. These were subclassifications of data under the general class, distribution of crimes. The schedule was drawn

up accordingly. The second step was collection of the required data. The third step was punching the information on machine cards and then assembling it on work sheets. The fourth step was tabulation.

Four general classifications are used in social statistics: chronological, geographical, magnitudinal, and qualitative. In Chapter III a dichotomous division of all statistical data was made, namely, attributes and variables. Chronological classes are usually variables, but not always so. Geographical classes are frequently not variables, but are attributes determined by political considerations. Magnitude classes are always variables. Qualitative classes are never variables; they are attributes, the definitions of which may be sufficiently precise for the profitable application of statistical methods of analysis. Of course, any number of subclassifications of the four main classifications mentioned above may be made; the number will depend upon the purpose in the mind of the investigator. The important point here is that the process of classifying the data will be almost complete long before the stage of tabulation is reached. If the data are in sufficient detail, they may be recombined in various ways to give new classes at the time of tabulation, but this too precedes the construction of tables, though it may come after the collection and assembling of the data.

2. CONSTRUCTION OF TABLES

A table is drawn on a flat surface, generally rectangular in form, ruled according to the requirements of the data. But certain steps are to be taken before the ruling is done. The worker must decide what *captions* are necessary and what is to be represented in the *stub* of the table. But care should be taken in thinking out the captions and stubs so as not to make the table too elaborate. A table is, after all, designed to simplify and summarize, and this purpose is defeated when it becomes too complex. This point can be discussed best from the table below for purposes of clarity. The *captions* are the headings in the spaces at the top of the table; they indicate the nature of the data contained in the columns. The "Year" and "Leather and Its Products" are the major captions, and they are coördinate. "Group Index," "Leather," and "Boots and Shoes" are captions subordinate to "Leather and Its Products." That is, they are subdivisions of the major caption, but they are coördinate with respect to each other. "Employment" and "Payroll Totals" are captions subordinate to the subdivisions of

TABLE III

INDEXES OF EMPLOYMENT AND PAY-ROLL TOTALS IN MANUFACTURING INDUSTRIES CONCERNED WITH LEATHER AND ITS PRODUCTS, YEARLY AVERAGES, 1923 TO 1929[1]

Year	Leather and Its Products					
	Group Index		Leather		Boots and Shoes	
	Employment	Pay-roll Totals	Employment	Pay-roll Totals	Employment	Pay-roll Totals
1923	110.7	113.9	109.6	107.0	111.1	117.0
1924	100.3	100.6	96.9	95.7	101.6	102.8
1925	101.9	101.8	98.7	97.5	102.9	103.6
1926	100.0	100.0	100.0	100.0	100.0	100.0
1927	97.9	97.4	98.4	97.2	97.7	97.6
1928	92.8	89.7	95.4	93.7	91.9	88.0
1929	92.8	89.9	92.2	93.2	92.9	89.0

[1] *Monthly Labor Review*, Vol. 30, No. 2, p. 186. The data here reproduced are taken from a larger table.

"Leather and Its Products." The *stub*, or the first column in the table, gives the second variable: time. "The units in which the measurements are made," says Secrist, "generally, although not always, appear in the 'caption': that is, in the vertical classes. The ways in which the measurements are presented generally, although not always, appear in the 'stub'—the horizontal classes. A tabulated datum, therefore, is found at the intersection of the vertical and horizontal axes."[1] A table, therefore, has two dimensions: vertical and horizontal. The characteristics of the data are given in the vertical dimension, or the columns, and the point of view from which they are regarded in the horizontal dimension, or the rows. A table thus in some degree presents an analysis of the data. Too much care cannot be given to the determination of captions and their relations of coördination and subordination; the clarity of the table depends upon this process.

Another point to be kept in mind is that coördinate captions may be both general and specific. "Group Index" is a general caption covering employment and pay-rolls in all leather and leather goods factories, but the coördinate captions, "Leather" and "Boots and Shoes," are specific; they are included in the group index but are separated for detailed analysis and presentation. When this kind of tabulation is necessary, the general class of data should be given in the first column to the right of the stub; the specific data are then in columns to the right of the general class.

[1] Secrist, Horace, *An Introduction to Statistical Methods*, pp. 128, 129. New York: Macmillan, revised edition, 1929.

This is a matter of convenience for two reasons: first, any reader will be interested in getting a general picture of the problem, before he goes to details; second, probably more people are interested in the general aspect alone than in both general and specific aspects. Furthermore, as a technical matter, it is the accepted mode of tabulation among statisticians.

A similar procedure is observed, when the totals of columns are published. The following table shows this fact:

TABLE IV

POOR ASYLUM INMATES CLASSIFIED BY AGE AND SEX, AUGUST 31, 1929. INDIANA [1]

Age Group	Both Sexes	Male	Female
All Ages	4,156	2,904	1,252
Under 3 years	6	4	2
3 and under 17	9	4	5
17 and under 30	75	35	40
30 and under 45	317	178	139
45 and under 60	832	556	276
60 and under 75	1,616	1,195	421
75 and over	1,264	908	356
Age not given	37	24	13

[1] Arranged from data in the *Indiana Bulletin of Charities and Corrections*, No. 182, p. 302.

The totals are given for "Both Sexes" and for "Male" and "Female" at the top of the table. This enables a reader to see at a glance the number of persons who were given care in the poor asylums, which is often the only fact wanted by a reader. This table exhibits again the general caption with specific captions which are placed to the right of the general caption.

The title of the table is important. It should be brief but should indicate the main facts given. It is not necessary that the title be a complete sentence; few titles of tables in standard statistical publications are complete sentences. The title of Table III does not attempt to give the details presented in all columns: it gives the general characteristics of the subject, namely, indexes of employment and pay-rolls in leather and leather-products industries; and the viewpoint from which they are presented, namely, the years 1923 to 1929.

The ruling of the table is determined to a large extent by the relations of the various captions. If vertical lines are used, they are dropped from the horizontal line which underlines a more general

caption. The line which separates the stub from the columns to the right is dropped from the topmost horizontal line. The topmost horizontal line is either one heavy line, or a double line. Some authors draw a double horizontal line between the lowest caption and the data in the columns. If this is done in tables which have totals, as Table IV, the double line is below the row of totals. The best practice regarding the bottom of the table is to draw either a single or a double horizontal line. If the bottom is left without a line, the table has the appearance of incompleteness. The ends of the table are generally left open, though some authors prefer to enclose the whole table by using end lines.

Footnotes to tables are generally placed in small type immediately below the table. They may be placed at the bottom of the page, but it is more convenient to place them nearer the table. The footnote may be only for the purpose of giving credit to the source from which the data are taken, or it may be to explain some unusual variation in the data which might escape the reader or might even be impossible for him to discover from the table at all. Everything about the table should be perfectly clear to the reader without the necessity of his debating in his mind the meaning of the author.

Tables may be classified as to whether they serve a general or a specific purpose. This distinction is important, when the worker prepares his table, because the users of the two types of tables are not the same. Discussing this subject, Mudgett says, "The descriptive terms used indicate the difference between the two types, the general-purpose table being designed as a repository of the tabulations in full detail; whereas the analysis table [or special-purpose table] is intended, as the name suggests, to present the results of analysis, to give not necessarily or always full detail, but summaries or conclusions and significant relationships."[2] The tables in the decennial publications of the United States Census are general-purpose tables. They are intended for thousands of persons whose interests in them vary widely. Public health statisticians want to know the details of age distribution so that they can calculate specific birth and death rates. Business men want to know the changing population by geographical areas so that they can estimate the future of their business in different parts of the country. Educators and social workers want to know the details about

[2] Mudgett, Bruce D., *Statistical Tables and Graphs*, p. 30. Boston: Houghton Mifflin Co., 1930.

school attendance and child labor. Students of population want to know the details of age groups by sex and the division into rural and urban population. The detailed data of the census tables may be rearranged into less detailed groupings, but if published in large groupings they could not be broken down into details. Many statistical reports of federal, state, and city departments publish general-purpose tables so that their data may be of the widest possible use. The special-purpose table may give only averages, percentages, or coefficients of correlation, or it may give the original data in frequency classes suitable to the purpose in hand, but too general for the use of many other workers. The special-purpose table, as Mudgett suggests, presents the results of analysis and conclusions.

Many people dislike to read a book or an article containing statistical tables. It appears to them formidable. For popular purposes the book or article without statistics has its place, but for the part of the public interested in knowing the facts about a subject statistical tables are essential. They enable the statistician to present his findings in brief space. How many pages of text would it take to present the facts brought out in Table III above? It would take quite a number, and, when the text was written, the reader would not have as clear an idea of the facts as he can now get in a few minutes' study. The table is indispensable for the presentation of masses of data, and the student should become accustomed to reading tables as a matter of course, and he should learn to think of his own data in terms of tables.

3. THE ARRAY

As data appear on a work sheet, they are unorganized. The student can have no idea of their meaning, until they are arranged in some orderly manner. Likewise published data may have a direct bearing upon a problem, but may not be in the order required for the purpose in hand. They must be reorganized to satisfy the requirements of the problem under consideration. Table V gives the number of jail prisoners per 100,000 population in Indiana, October 1, 1928, to September 30, 1929, according to counties. It is obvious that the arrangement of counties in alphabetical order has no significance in so far as the occurrence of imprisonment in jails is concerned. No conception of the average rate of such imprisonments can be obtained from this table.

TABLE V

JAIL PRISONERS PER 100,000 POPULATION IN INDIANA BY COUNTIES, OCTOBER 1, 1928, TO SEPTEMBER 30, 1929 [1]

County	Prisoners per 100,000 Pop.	County	Prisoners per 100,000 Pop.
Adams	389	Lawrence	2,348
Allen	812	Madison	1,412
Bartholomew	1,626	Marion	1,614
Benton	493	Marshall	600
Blackford	1,044	Martin	850
Boone	1,690	Miami	1,955
Brown	1,856	Monroe	2,637
Carroll	543	Montgomery	1,879
Cass	1,817	Morgan	1,347
Clark	2,108	Newton	893
Clay	622	Noble	388
Clinton	902	Ohio	1,968
Crawford	358	Orange	1,070
Daviess	1,052	Owen	952
Dearborn	1,530	Parke	836
Decatur	999	Perry	887
Dekalb	689	Pike	623
Delaware	2,180	Porter	1,253
Dubois	285	Posey	1,092
Elkhart	487	Pulaski	1,344
Fayette	2,525	Putnam	2,010
Floyd	1,551	Randolph	878
Fountain	708	Ripley	324
Franklin	56	Rush	1,061
Fulton	——	St. Joseph	1,256
Gibson	744	Scott	894
Grant	2,275	Shelby	1,091
Greene	561	Spencer	665
Hamilton	948	Starke	904
Hancock	3,637	Steuben	981
Harrison	433	Sullivan	1,497
Hendricks	978	Switzerland	497
Henry	2,477	Tippecanoe	1,451
Howard	1,547	Tipton	784
Huntington	935	Union	1,494
Jackson	515	Vanderburgh	51
Jasper	1,215	Vermilion	1,113
Jay	475	Vigo	3,826
Jefferson	843	Wabash	871
Jennings	368	Warren	793
Johnson	1,270	Warrick	814
Knox	1,533	Washington	831
Kosciusko	503	Wayne	1,515
Lagrange	634	Wells	513
Lake	1,158	White	383
Laporte	792	Whitley	659

[1] Rates computed from data of *Indiana Bulletin of Charities and Corrections*, No. 182, pp. 307, 308.

The simplest form of orderly arrangement would be in the form of an array, that is, listing them in order of magnitude from lowest to highest. Table VI presents the jail rates as an array with the names of the counties omitted:

TABLE VI

JAIL PRISONERS PER 100,000 POPULATION IN EACH COUNTY OF INDIANA, OCTOBER 1, 1928, TO SEPTEMBER 30, 1929, ARRAYED ACCORDING TO RATE

Jail Prisoners per 100,000 Population			
51	659	952	1,530
56	665	978	1,533
285	689	981	1,547
324	708	999	1,551
358	744	1,044	1,614
368	784	1,052	1,626
383	792	1,061	1,690
388	793	1,070	1,817
389	812	1,091	1,856
433	814	1,092	1,879
475	831	1,113	1,955
479	836	1,158	1,968
487	843	1,215	2,010
493	850	1,253	2,108
503	871	1,256	2,180
513	878	1,270	2,275
515	887	1,344	2,348
543	893	1,347	2,477
561	894	1,412	2,525
600	902	1,451	2,637
622	904	1,494	3,637
623	935	1,497	3,826
634	948	1,515	

From the array it is easy to see that the rates below 1,000 predominate and that there are few counties with rates of over 2,000. Two extremely low rates and two extremely high rates appear. The two lowest rates are proportionately so much lower than the next highest that it is probable some extraneous factor in recording and reporting is responsible for the difference. The two highest rates are not so much higher than the rates just below them to appear impossible. The array shows up still better in graphic form. Figure XVI presents the above data graphically.

Examination of this figure reveals the wide range from the lowest to the highest rates of jail imprisonment. Possible explanations of the large differences are many: (1) there may be real differences in the tendency to crime and delinquency in various counties; (2) there may be wide differences in the strictness with which the

law is enforced; (3) some communities may permit bail more easily than others; (4) differences in reporting jail imprisonments may account for some differences. It is obvious that the array of imprisonment rates of counties does not explain why differences occur but simply makes clear that they exist. One of the functions

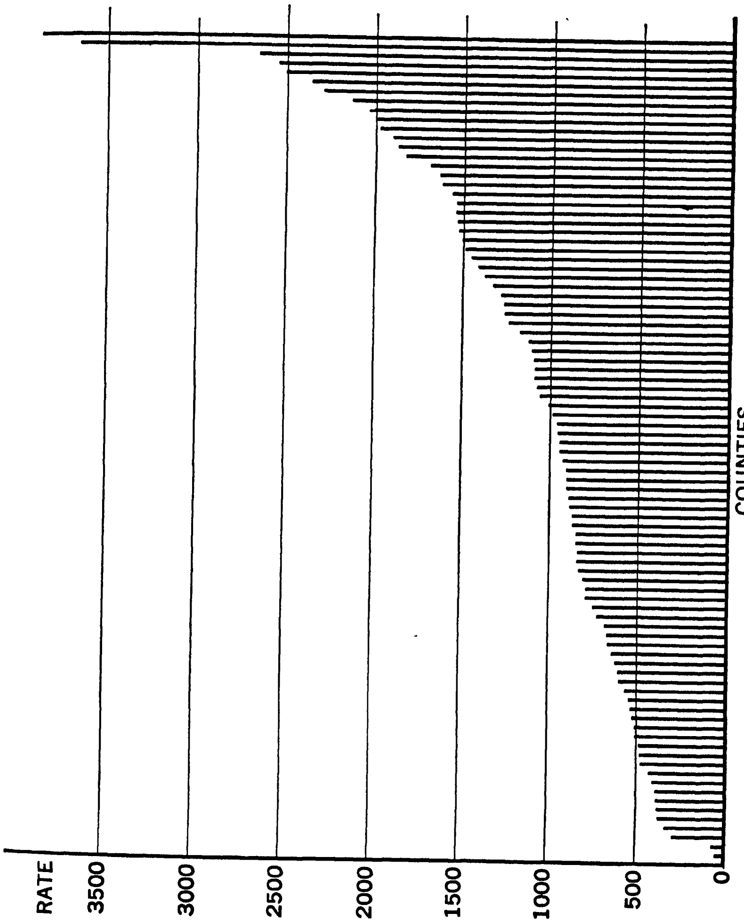


FIGURE XVI.—JAIL PRISONERS PER 100,000 POPULATION IN INDIANA COUNTIES

of statistics is to reveal similarities and differences in masses of data.

The reader will be aware of questions to which he would like to have answers, which statistics can answer but which are not answered by the array alone. Around what rate of imprisonment

do the rates tend to cluster? If the array is divided into parts with equal ranges, in what part do the largest number of rates appear? The array cannot answer such questions. That is the function of the frequency distribution, to which we shall now turn.

4. THE FREQUENCY DISTRIBUTION

The frequency distribution is defined by Chaddock as follows: "An arrangement of quantitative data in order of magnitude, grouped by a selected class-interval of value so as to reveal clearly the internal structure of the mass of facts for the purpose in view, and so as to be accurate and useful for purposes of summarization, comparison, and analysis."[3] If the frequency distribution is to do all this, that is, reveal the internal structure of the data and be useful for summarization, comparison, and analysis, it must be carefully constructed. It is a fundamental process in statistical analysis. Table VII presents the data of Table VI in the form of a frequency distribution:

TABLE VII

FREQUENCY DISTRIBUTION OF JAIL IMPRISONMENT RATES ACCORDING TO COUNTIES

Rate	Number of Counties
All	91
Under 500	14
500–999	36
1,000–1,499	18
1,500–1,999	12
2,000–2,499	7
2,500–2,999	2
3,000–3,499	0
3,500–3,999	2

The concentration is in the range from 500 to 999; more than a third of all the counties have these rates, and less than half have rates greater than 999. In view of this fact it would be interesting to know why a small number of counties have rates much greater than the lower half, but all these statistics can do is to raise this question.

It will be noticed in Table VII that the rates are grouped in intervals of 500. All counties with rates less than 500 are put in the class-interval, 0-500; all the counties with rates of 500 but

[3] *Op. cit.*, p. 57.

less than 1,000 are put in the class-interval, 500-999; etc., etc. The number of counties whose rates fall within the limits of a class-interval is known as the class-frequency. In view of the fact that the frequency table is intended to convey some idea of the central tendency, or average magnitude, of the data, the size of the class-interval becomes important. Looking at the second class-interval of the table and noting that 36 counties have rates between 500 and 999, one almost automatically thinks of the average rate of this class-interval as 750, that is, the mid-point of the class-interval. In an even distribution that is a fact, and it is the assumption made in dealing with all frequency distributions. Therefore, it is important to select a class-interval most accurately representing the data. For example, the simple arithmetic average of the rates in Table VII, found by adding all the rates and dividing by 91, is 1,118; that is the absolute arithmetic average. When the arithmetic average is computed from the data grouped by class-intervals of 250, it is found to be 1,158; by class-intervals of 500, it is 1,129; by class-intervals of 1,000, it is 1,094. The average closest to the absolute average is that computed from the data arranged in class-intervals of 500. If the number of counties were large, say, 1,000 or more, the average computed from grouped data should be approximately the same as the simple average found by adding all items and dividing by the number of items. But even when the number of items is large, the size of the class-interval is important. In the class-interval, 2,500-2,999, there are only two items. Both are less than 2,750, but as a matter of fact they are assumed to be 2,750 in using the grouped data. The effect is to raise their value, and, hence, it is not surprising that the average computed with a class-interval of 500 is slightly larger than the simple average. It might just as well be smaller than the simple average, as is the case when computed from data grouped by class-intervals of 1,000.

This raises the question of artificial concentration at certain values in a frequency distribution. Table VIII makes this point clear.

Notice the concentration of frequencies on grades divisible by 5. In grading papers an exact evaluation of the work is generally impossible. Since people, including teachers, generally think more easily in terms of numbers divisible by 5, grades tend to be given in this manner. If it were decided to put the above data into a frequency distribution with class-intervals greater than 1, the mid-

TABLE VIII

FIVE HUNDRED MARKS IN ENGLISH CLASSIFIED BY SINGLE PER CENTS[1]

Grade Per Cent	Frequency	Grade Per Cent	Frequency
20	20	52	10
21	0	53	3
22	1	54	3
23	1	55	20
24	0	56	0
25	20	57	1
26	0	58	4
27	0	59	0
28	0	60	25
29	1	61	3
30	38	62	13
31	0	63	8
32	3	64	2
33	3	65	15
34	3	66	0
35	47	67	2
36	1	68	6
37	0	69	0
38	9	70	19
39	2	71	1
40	53	72	2
41	0	73	0
42	4	74	0
43	2	75	10
44	2	76	0
45	55	77	1
46	0	78	1
47	5	79	0
48	18	80	7
49	0	85	3
50	46	90	3
51	4		

[1] Data from Chaddock, *op. cit.*, p. 77.

point of the class-interval should in all cases be a number divisible by 5. Table IX presents these data in class-intervals of 5.

The arithmetic average of the grades arranged in intervals of 5 with the numbers divisible by 5 falling at the mid-point is 47. If the class-intervals are left the same size but rearranged so that the numbers divisible by 5 fall at the top of each class-interval, the average is 45.5. That is not a great difference, but it illustrates the effect of the class-interval upon the average. The student will frequently find data which for some artificial reason tend to concentrate at numbers divisible by 5, 10, 25, 50, 100, 500, 1,000. Salaries are likely to be in terms of hundreds of dollars. If they are classified into a frequency distribution, the mid-point of the

TABLE IX

FIVE HUNDRED MARKS IN ENGLISH CLASSIFIED BY INTERVALS OF 5 PER CENT

Grade Per Cent	Frequency
18–22	21
23–27	21
28–32	42
33–37	54
38–42	68
43–47	64
48–52	78
53–57	27
58–62	45
63–67	27
68–72	28
73–77	11
78–82	8
83–87	3
88–93	3

class-interval should fall on an even 100 or 1,000. There is often seen some concentration of ages around numbers divisible by 5. Retail prices of articles fall more often on numbers divisible by 5 than on any other. Likewise wages are likely to be on even dollars, half-dollars, or quarters, though piece wages are more evenly distributed. Whenever there is any reason to suspect an artificial factor operating to bring about concentration around certain numbers, these numbers should be ascertained before the class-interval is decided upon, and then, if these numbers recur regularly, they should be placed at the mid-point of the class-interval.

Two other considerations enter into determining the size of the class-interval. General-purpose tables should have small class-intervals—intervals as small as anybody is likely to want. Special-purpose tables may have class-intervals of any size that gives satisfactory results. Such data as those published by the Bureau of the Census are for general use. The age distribution must be given in small class-intervals so that they may be used by persons who want a single-year distribution as well as by those who want 5- or 10-year distributions. The larger class-intervals can be made up from the small ones, but the large ones could not be broken down into the small ones. For many purposes it is desirable to know the number of the population for each year of age, especially below five years of age. The census reports give these numbers, though they generally give the total for the 5-year period also. If the

statistician has assembled a large mass of data for a special purpose but has an idea that he might use it for some other purpose, he must keep the data on file in class-intervals as small as he would ever want, though he may publish the results of a special study and use only large class-intervals.

Occasionally a table does not have class-intervals of uniform size. Small class-intervals are used for the lower magnitudes, but large ones are introduced for presenting the frequencies in the higher magnitudes. This is sometimes done, because the frequencies in the larger magnitudes are small in number. For example, in Table VII only 11 counties have imprisonment rates of 2,000 or more. All of these might have been grouped in a class-interval of 2,000-3,999. An average computed from such a grouping would likely vary considerably from the true average. With such a grouping in Table VII the average would be 1,186. This is much larger than the true average. Thus, it will be seen that, if the grouped data are to be used for obtaining an arithmetic average, they should be presented in uniform class-intervals. If there are special reasons for using class-intervals of different sizes in the same table, then the larger ones should be multiples of the smallest class-interval used. For example, the smallest class-interval might be 5, as in Table IX, but in the upper ranges the class-interval might be increased to 10 or 15. But even this practice limits the uses to which someone else might want to put the data. In special-purpose tables there is more justification for class-intervals of different sizes, but there is hardly any justification for the practice in general-purpose tables.

There is a device which may be used with approximate accuracy to redistribute class-frequencies, if they happen to be given in class-intervals unsuitable to the purpose of the worker. This is a cumulative frequency curve.[4] Suppose we have the census distribution of population in a city by age-groups and for some special purpose we need a different distribution. How could we determine the number of children 11 to 13 years of age, if we have only the number for 10 to 14 years of age given in the table? A cumulative frequency curve with age as the horizontal scale and numbers of the population for the vertical scale can be made. Then the number indicated by the curve at 13 years is found and the number indicated at 11 years is found. If we subtract the second

[4] See p. 175ff. for detailed description of cumulative frequency curves.

number from the first, we have approximately the number of children 11-13 years of age.[5]

The limits of the class-interval should be determined and stated precisely. In Table VII the first class-interval is given as "under 500." That means that any rate falling short of 500 by however small an amount is placed in this class-interval, and the assumption is that in each of the other class-intervals rates falling short of the lower limit of the next class-interval by however small an amount belong in the class-interval below this limit. There is, then, no question as to what rates belong and are put in each class-interval. But suppose that the first class-interval were written "0-500" and the next one "500-1,000." Where would a rate of 500 be put? Only the person who constructed the table could tell, and he might have forgotten just what he did. The class-intervals should be stated in numbers which are mutually exclusive.

5. EXERCISES

1. The forms which freshmen fill out at college, when they matriculate, are a good source of data for practice in constructing tables. These data are already gathered and require no field work on the part of the student. From them construct the following tables:
   (a) Age distribution of freshmen by sex.
   (b) Credits offered by freshmen to meet admission requirements. Make a frequency table.
   (c) Occupations of the parents of freshmen.
   (d) Height and weight.
2. Construct a schedule suitable to obtain the following information from students: age, sex, occupation of parents, occupational intentions of the student, height, weight, nationality, race. Each student should take a number of these schedules and get the necessary information from his friends. If no names are taken, no objections should be encountered. The information obtained by all the students can then be pooled so that each one will have sufficient data with which to work. Construct tables which exhibit the meaning of the data.
3. Take 100 leaves from a tree, measure the length of each leaf, and present the measurements graphically as an array.
4. The following data are miles which 415 male felons in Indian-

[5] This device is illustrated by Whipple, G. C., *Vital Statistics*, pp. 75-77. New York: Wiley, 1923.

apolis in 1930 went from their homes to commit the offenses for which they were sentenced by the court. Make frequency tables from these data, using class-intervals of half a mile and one mile:

.86	.95	3.70	3.30
1.30	1.00	.54	1.86
1.00	2.81	2.16	1.00
2.54	4.41	1.76	.76
.89	2.11	2.13	2.13
4.46	2.89	.38	.97
3.19	1.73	.76	4.08
1.24	.62	2.22	1.51
8.43	.95	.95	4.21
2.08	2.02	4.05	1.81
1.30	.62	4.30	4.24
1.08	5.62	3.89	4.11
.76	8.43	3.35	3.03
3.52	2.27	.86	.54
1.00	1.00	2.00	3.76
2.76	1.05	1.05	3.73
2.89	3.16	.76	.92
	1.08	.49	2.37
2.16	2.00	2.11	2.00
2.89	.87	2.37	.76
.54	1.24	3.25	.49
4.43	1.08	.95	2.00
2.22	1.11	.49	2.16
3.14	2.68	3.00	2.62
38	1.41	7.89	2.87
03	2.00	.95	1.62
95	3.76	1.03	1.08
92	.39	.81	.97
62	1.62	4.43	3.03
.54	.97	1.00	1.49
2.30	.38	2.89	6.59
7.41	2.81	1.84	2.70
1.57	.95	.86	1.08
1.41	1.92	1.14	4.76
1.49	.27	.54	1.97
2.49	2.76	1.19	1.08
.62	.76	.68	.97
54	.49	3.51	.76
2.16	1.00	1.03	1.22
4.22	6.68	1.19	.76
1.03	1.68	2.14	1.30
.95	8.35	4.76	5.03
1.16	1.87	4.57	3.08
1.43	3.68	.95	5.16
1.41	1.16	3.30	1.49
4.30	4.08	1.19	.46
.86	3.14	2.38	
1.70	.70	2.05	1.41
.73	.00	1.11	2.92
1.41	.59	1.22	1.65
2.65	.30	3.92	.97
1.00	.41	4.14	2.24
2.08	.76	1.51	3.65
2.05	.57	2.35	3.01
2.81	3.01	.97	2.38
3.43	2.16	1.54	2.22
1.08	1.27	4.14	2.00

1.49	1.16	.51	.78
2.32	2.97	1.35	2.11
1.76	3.87	1.03	.73
2.27	.51	.81	1.46
1.86	1.16	.51	3.43
.92	.95	1.05	.95
.95	1.16	.95	1.97
1.97	.84	1.65	2.54
3.38	.81	2.03	1.65
3.95	2.11	1.86	2.08
1.16	4.87	1.11	1.11
1.19	1.19	3.95	1.57
.86	2.27	4.65	.51
.92	1.00	.86	1.16
.97	4.33	3.03	1.00
2.00	.81	.81	2.21
1.14	4.05	4.65	.95
2.32	1.87	.51	1.03
1.92	.86	1.00	.54
2.76	.81	1.97	1.27
.73	6.16	3.32	1.97
3.68	3.35	.73	2.54
1.51	.46	1.30	4.00
2.00	1.73	3.46	3.11
1.51	2.51	2.16	3.24
1.08	2.81	1.19	1.24
2.03	1.19	.57	2.22
1.62	.46	1.32	.86
.76	1.05	.97	3.83
2.24	.57	3.24	2.16
1.30	1.30	.57	.57
1.22	1.19	1.78	1.78
1.84	.78	1.19	.43
1.35	2.22	2.03	2.24
1.30	1.11	.62	2.22
1.73	2.89	1.14	4.59
2.03	6.16	2.46	1.19
2.79	2.79	2.06	2.65
1.24	1.00	2.06	1.00
2.81	2.33	.49	.49
1.00	5.68	1.08	1.38
.86	2.65	1.19	.38
.46	.92	.86	1.68
2.75	.65	1.76	5.19
.38	2.75	1.78	3.30
1.43	2.03	1.68	1.51
4.60	1.73	.97	

6. REFERENCES

Burgess, R. W., *Introduction to the Mathematics of Statistics*, Chap. IV.

Chaddock, R. E., *Principles and Methods of Statistics*, Chap. V.

Gavett, G. I., *First Course in Statistical Method*, Chap. II.

Mills, F. C., *Statistical Methods*, Chap. III.

Mudgett, B. D., *Statistical Tables and Graphs*, Part I, Chap. III.

Secrist, Horace, *An Introduction to Statistical Methods*, Chap. VI.

Yule, G. U., *An Introduction to the Theory of Statistics*, Chap. VI.

CHAPTER VII

Graphic Presentation

I. INTRODUCTION

GRAPHIC presentation of social statistics is a way of making abstract relations and magnitudes visible by means of symbols. A graph appeals to the eye. It pictures relationships and magnitudes in various symbols having conventionally accepted meanings. Graphic methods are introduced fairly late in the study of a social problem involving statistics. Long before they are required, the problem has been defined and data have been collected, tabulated, and classified. Even after the data have been classified, some other statistical analysis may be undertaken before graphs are constructed. But the analysis done at this point is more likely than not to involve the use of graphic methods. Graphic methods are in many respects simple, but it will be seen in this and later chapters that line graphs may become rather complex in conception. Thus, it will be seen that graphic methods serve an analytical as well as a presentational purpose. This chapter is concerned with graphs of

TABLE X

THE NUMBER OF NEW PROTESTANT DENOMINATIONS IN EACH 50-YEAR PERIOD, 1500 TO 1900, AS REPRESENTED IN THE UNITED STATES [1]

Period of Origin	Number of New Denominations in Each Period
1500–1549	4
1550–1599	2
1600–1649	7
1650–1699	3
1700–1749	6
1750–1799	10
1800–1849	43
1850–1899	80

[1] See White, R. Clyde. *Denominationalism in Certain Rural Communities in Texas*, p. 12. Training Course for Social Work, Indiana University, Indianapolis, 1928.

both kinds, though for exhaustive treatment the reader is referred to standard monographs on the subject of graphs. The presentational and analytical functions of graphs cannot be entirely separated. Sometimes a graph which announces certain facts in an emphatic way also serves an analytical purpose, and vice versa. This double function of graphic methods is illustrated below:

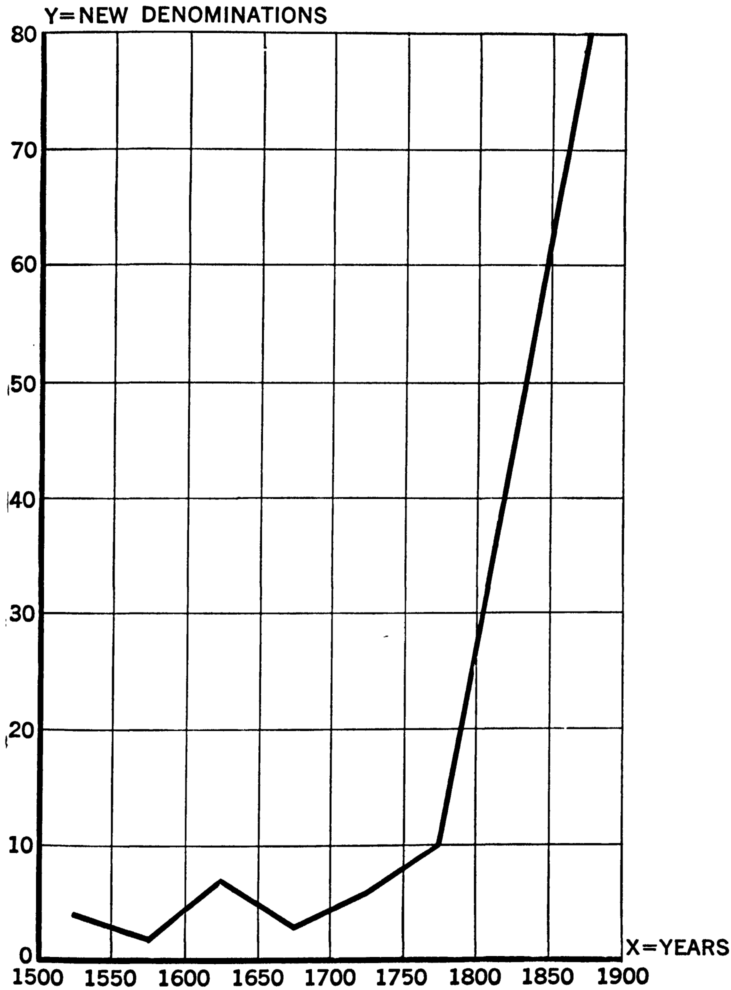


FIGURE XVII.—NEW PROTESTANT DENOMINATIONS IN EACH 50-YEAR PERIOD, 1500 TO 1900, AS REPRESENTED IN THE UNITED STATES

This chart shows that the tendency of the Christian Church to split into denominations or sects has been greater in recent times than

in the period immediately following the Protestant Reformation. Any person glancing at the title of the chart, at the base line and left vertical line designations, and then at the curve would infer that new denominations arose much more rapidly in the second half of the nineteenth century than in any previous fifty-year period. The distance of points from the base line measures the rapidity of increase in denominations. As a method of analysis the chart shows change in number of denominations by definite periods. Table X, of course, gives the same result. But psychologically there is little doubt that the graph is more effective in convincing the reader of the strength of the drift to denominations. It presents the facts in their correct relations, and presents them effectively. It would do this without the table, but it is better to give the table also so that anyone who wishes may consult the exact figures.

So much for what the chart shows. But the student beginning the study of statistics is interested in the mechanics of the graph. Periods of time are represented on the base line. One side of a square represents each period of fifty years, beginning at the left and going forward with time toward the right. In any graph in which time is one of the factors to be plotted, whether days, months, or years, it is customary to plot time along the base line. The other factor is plotted on the vertical line to the left, as indicated on this chart. The vertical scale starts with zero at the bottom and goes as high as the data require. Another thing to notice is that the points representing time are located in the middle of the squares in the horizontal direction. This is customary, because a definite period of time has elapsed, and it is assumed that some denominations arose early in each fifty-year period, some about the middle, and some toward the end. Placing the point halfway between 1550 and 1600, or any other two terminal dates, gives each end of the period equal weight.

In Chapter III it was pointed out that there are independent and dependent variables and that the statistician is primarily concerned with the relations existing between them. Time is always an independent variable. Whatever social phenomena appear, they must appear in time and in a definitely measurable period of time. Hence, the fifty-year periods of time in Figure XVII are the independent variable. The independent variable is by convention plotted on the horizontal line and is designated by X. In this problem "new denominations" are the dependent variable. They

occur in time, and they cannot occur without the passage of time. They have become more frequent as time has passed. The variation in number of denominations is a function of time. No arbitrary values can be assigned to "new denominations"; they are dependent upon the operation of other factors which are not measured here—only time in which the variations occur is measured. The frequency of the variations in "new denominations" is caused by other factors. The dependent variable is plotted on the vertical line and is designated Y.

The type of graph represented by Figure XVII is known as a line graph, because the data are represented by points connected by straight lines, or they might be represented by a smooth line drawn to fit the distribution of points. The most common line graphs are straight line graphs, nonlinear graphs, ratio charts, histograms, and frequency polygons. All these types of curves will be found to fit various kinds of social data.

Line graphs are particularly useful in showing functional relationships; that is, the relations of two series of data, or variables, which are causally related. Other graphic forms are bar charts, pie charts, pictograms, and cartograms; these will be discussed briefly under the heading of "Miscellaneous Graphic Devices" in the latter part of the chapter.

Before proceeding to the detailed consideration of line graphs, two principles of general usefulness should be described: rectangular coördinates and logarithms.

2. RECTANGULAR COÖRDINATES

The principle of rectangular coördinates is involved in the construction of all line graphs. It sounds like a formidable mathematical concept, but in fact it is an elemental fact of common experience, though we do not ordinarily think of Cartesian coördinates when this experience comes along. Suppose a man plans to build a house on a rectangular lot. He wants to place the house accurately. The lot is 100′x125′, and the house is to have a 40-foot front. He decides that the house should be 30 feet from the south side of the lot and that the west side of the house should be 30 feet from the west side of the lot. If he measures off 30 feet directly east from the southwest corner of the lot and then turns north and measures off 30 feet, he will have located the point at which the southwest corner of the house will fall. In the following chart P indicates the southwest corner of the house:

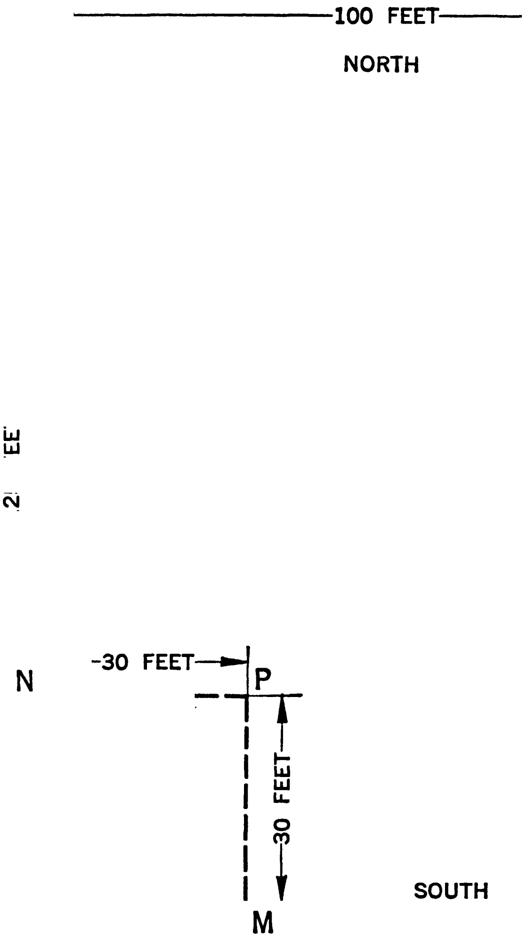


FIGURE XVIII.—LOCATION OF THE SOUTHWEST CORNER OF THE HOUSE AT P

Referring to the chart, the line MP is erected perpendicular to the south side at a point 30 feet from the corner, and the line NP is drawn perpendicular to the west side, which, of course, intersects the west side at a point 30 feet above the corner. The intersection of the lines MP and NP determines the location of the southwest corner of the house; these lines are the coördinates of the point P. Since these two lines intersect at right angles to each other, they are rectangular coördinates. Thus, such a common experience as locating the corner of the foundation of a house involves the principle of rectangular coördinates.

But how is this principle related to a line graph? The complete system of rectangular coördinates would be represented by four adjoining house lots, as in Figure XIX:

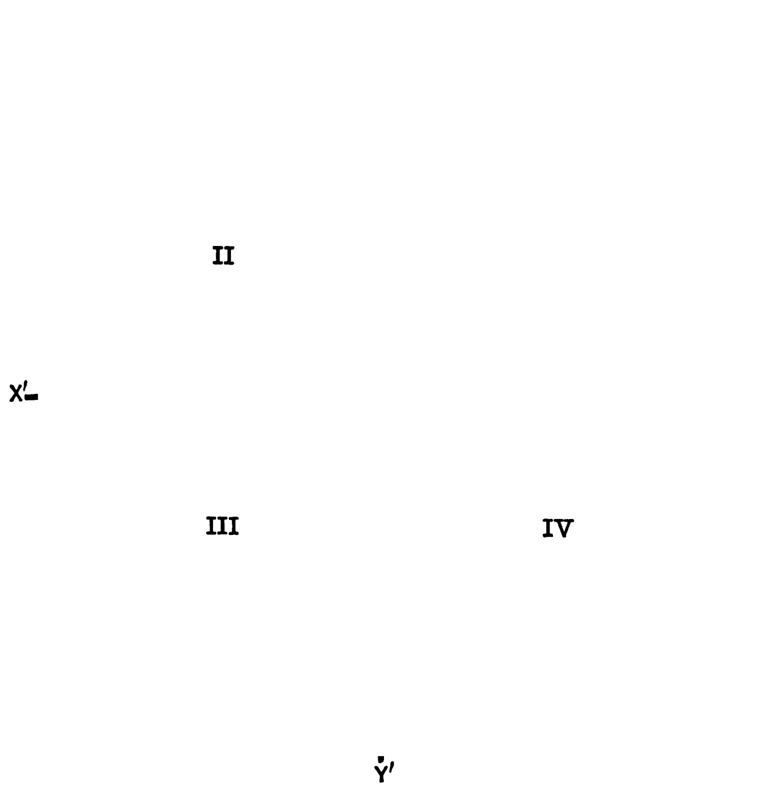


FIGURE XIX.—RECTANGULAR COÖRDINATES

The upper right section of the chart is known as the first quadrant, or the house lot represented in Figure XVIII, and the other quadrants, or house lots, are numbered II, III, and IV. But, dropping the analogy of the house lot, we have plotted the data from Table X in the first quadrant. Each point represents the intersection of coördinates, and the line connecting the points makes the curve.

Each of the coördinates of a point has a name. The horizontal coördinate is called the *abscissa*, and the vertical coördinate is called the *ordinate*. The base line is commonly designated X, and the vertical line on the left is designated Y. The point of inter-

section of these two perpendicular coördinates is designated O and is called the *origin*, or *zero origin*, meaning that both the X coördinate and the Y coördinate at this point have a value of zero. In plotting data for a curve the units on both the horizontal scale and the vertical scale are measured off from the origin.

One other conventional practice in the use of coördinates should be noted, and that is the positive or negative sign of the coördinates in different quadrants. Both the abscissa and the ordinate are positive in the first quadrant. In the second quadrant the ordinate is positive, but the abscissa is negative. Both coördinates are negative in the third quadrant. In the fourth quadrant the abscissa is positive, but the ordinate is negative. The general rule is that the abscissa is positive on the right of the origin and negative on the left of the origin. Correspondingly, the ordinate is positive above the origin and negative below the origin. In social statistics the first quadrant is used almost exclusively, though occasionally, as will be seen in Chapter XIII, the fourth quadrant will be used. It is conceivable that one might set up a statistical problem involving social data which would require the use of other quadrants. Graphs like Figure XVIII will be the more common, however, and only the first quadrant will appear in the presentation.

Two other definitions are necessary. Referring to Figure XIX the line XX′ is known as the x-axis, and the line YY′ is known as the y-axis. Instead of referring to the base line or the vertical line on the left, it will be convenient to speak of the x-axis and the y-axis.

3. LOGARITHMS

Logarithms have a variety of uses in statistical work, particularly in graphic presentation. They are used most frequently in calculating geometric averages, certain index numbers, and in logarithmic and semi-logarithmic curves. A brief account of the theory and use of logarithms is necessary at this point.

A logarithm is the power of a number, known as the base, to which the number must be raised to equal a second number. For example, 2 is the power to which 10, the base number, must be raised to equal 100, and 2 is the logarithm of 100. The power, 2, is the exponent of 10, that is, 10 is to be squared, and as the logarithm of 100 it represents the root of 100 which must be found in order to determine the base number. If b is the base,

x the power to which the base is to be raised, and N the number, the exponential form is

$$b^x = N$$

$$\text{or} \qquad 10^2 = 100$$

The logarithmic form is

$$x = \log_b N$$

$$\text{or} \qquad 2 = 2.000000$$

To use logarithms in multiplying two numbers the logarithms are added, and the sum of the logarithms of the numbers is equal to the logarithm of the product of the numbers. Then the number which is the product of the numbers may be found in a table of logarithms. Similarly, the logarithm of the quotient of two numbers is equal to the difference of the logarithms of the numbers, and the quotient of the numbers is found in a table of logarithms. The square root, the cube root, or any other root of a number may be found by dividing the logarithm of the number by the index (e.g., 2 for the square root) of the root required. The quotient of this operation is the logarithm of the number which is the root required. This is important to remember, because many students will have forgotten how to extract the square root of a number and most of them will not know how to extract higher roots, whereas the use of a table of logarithms for this purpose is simple. Many later problems in this text will require square roots.

The common system of logarithms is calculated on the base, 10. However, a system of logarithms might be calculated upon any number as the base. The decimal system is more convenient, and the published tables of logarithms all use this base. Appendix C contains logarithms for numbers from 1 to 11,000 true to five decimal places. If the logarithm of 100 is 2, then the logarithm of 1,000 is 3, since 10 raised to the third power is 1,000. What would be the logarithm of a number lying between 100 and 1,000? For example, 756. Consulting the table of logarithms, we find in the first column to the right of the number, 756, the number, 878522. The logarithm of 756 obviously will be between 2 and 3. This large figure found in the table should have a decimal point in front of it and, to the left of the decimal point, the number, 2. Hence, the logarithm of 756 is 2.87852.

There are two parts to every logarithm. That part to the left of the decimal point never appears in the table, because it is de-

termined from the number of digits in the number. This part of the logarithm is known as the *characteristic* and is always 1 less than the number of digits in the number—i.e., the number of digits which lie to the left of the decimal point, if there is one in the number. Therefore, the characteristic of the logarithm of 756 is 2. That part of the logarithm which is to the right of the decimal point is called the *mantissa*. This is the part of every logarithm which is found in the table. The mantissa of a number is always positive. The characteristic of a number greater than 1 is positive, but the characteristic of a number less than 1 is negative. Suppose it is desired to know the logarithm of .00289, a number which is less than 1. Look up the mantissa of 289 in the table—the mantissa of 289 is the same, whether the number be 289, 28.9, 2.89, or .00289. The mantissa is found to be .46090. The characteristic of a number less than 1 is negative and is 1 greater than the number of zeros between the decimal point and the first significant figure. Write the logarithm of .00289 thus: $\bar{3}$.46090, or 7.46090-10.

After examining the table of logarithms it will be noticed that there are 10 columns of figures and that at the top of each is a number in heavy-face type. These numbers run from 0 to 9. If the logarithm of 289 is required, it is found to be 2.46090. But suppose the number is 289.7. What is the logarithm? In the column with 7 at the top and opposite 289 is the number 46195. Supplying the characteristic, we have 2.36195 which is the logarithm of 289.7. If the number were 289.74, a slightly different problem is presented, because the exact logarithm for this number is not given but must be found by interpolation. We subtract the mantissa of the logarithm for 289.7 from the mantissa of the logarithm for 289.8 which gives a remainder of .00015. The significant figures in this quantity are 15. One of the little tables on the margin of the page has 15 in heavy-face type at the top. We run down the column of heavy-face type figures at the left until we get to 4 which is the digit at the extreme right in 289.74. Opposite this number in the table and in the next column is 6.0, or it is really .00006. If this number is added to 2.46195, the sum is 2.46201 which is the logarithm of 289.74.

To find the number which corresponds to a logarithm the above procedure is reversed. It should be remembered that only the mantissa can be found in the table. Suppose the logarithm is 2.46201. What is the number to which it corresponds? Turning to the table of logarithms, the first column of light-face type is

followed down until 46 is found. Then the remainder of the mantissa will be found in another column and possibly in a different row of mantissas. The nearest mantissa to .46201 is .46195. That is the mantissa of 289.7. It is not the logarithm which we have. Subtracting the mantissa, .46195, from the mantissa, .46201, in the next column we get 15. The difference between our mantissa and .46195 is .00006. Consulting the table of proportional parts which has 15 at the top, we follow down the column of light-face type until we find 6 or the number nearest to it. Opposite this number in the column to the left is 4. That is the last digit of the number sought, which is 289.74.

4. THE STRAIGHT LINE GRAPH

For some data in social statistics the graph is a straight line. This is due to the fact that the quantities change by equal increments or decrements in a specified period of time. Figure XX will illustrate this principle. The data used in this graph are drawn from the field of crime and are given in Table XI. If a man is sentenced to federal prison for 10 years, he may reduce his time at the rate of 10 days per month for good conduct.[1] That is, when he has served a year of 365 days, he may get credit for having served 486 days. The table and graph follow:

TABLE XI

THE ANNUAL ACCUMULATION OF THE PERCENTAGE OF A 10-YEAR SENTENCE SERVED BECAUSE OF GOOD CONDUCT IN A FEDERAL PRISON

Year	Percentage of Sentence Served, End of Each Year
First	13.28
Second	26.56
Third	39.84
Fourth	53.12
Fifth	66.40
Sixth	79.68
Seventh	92.96
Eighth (.53 yr.)	100.00

If a prisoner received no deductions from his sentence for good behavior, his sentence would be represented by the broken line (1), but, if he has a perfect record and received maximum deductions for good behavior, his time served would be represented by the

[1] *The Code of Laws of the United States of America*, p. 514, sec. 710. In force December 6, 1926.

solid line (2). For perfect conduct the percentage of his sentence served in a year of 365 days is not 10.0 per cent, but 10.0 per cent plus 3.28 per cent. At the end of each year 13.28 per cent of his total sentence would be deducted from what remained so that he would be released from prison soon after the middle of the eighth year instead of at the end of the tenth year.

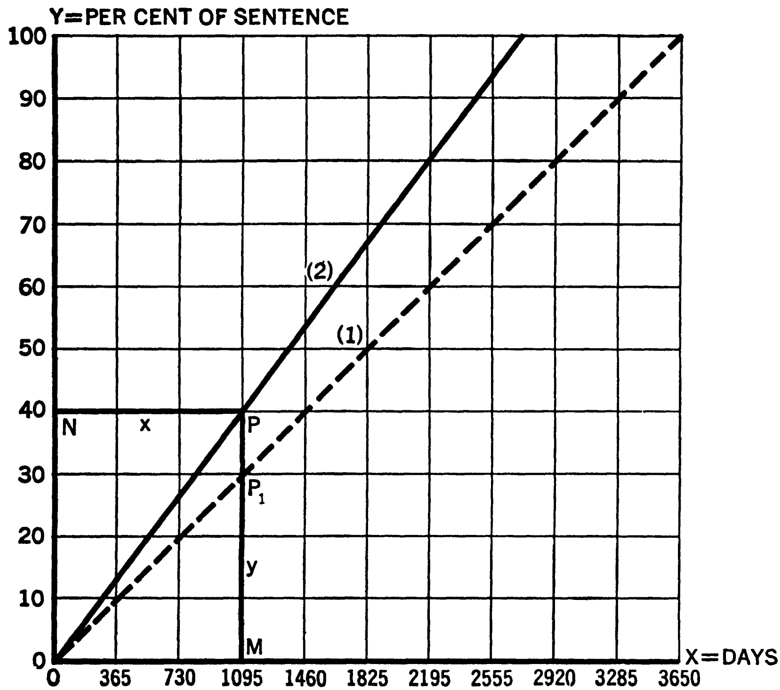


FIGURE XX.—SHOWING THE CUMULATIVE PERCENTAGE OF TIME SERVED ON A 10-YEAR SENTENCE IN A FEDERAL PRISON (1) WITHOUT DEDUCTIONS FOR GOOD BEHAVIOR AND (2) WITH REGULAR MONTHLY DEDUCTIONS FOR GOOD BEHAVIOR

Sometimes it is desirable to express a straight line in terms of an equation. This is particularly important if the slope of one straight line is to be compared precisely with the slope of another straight line. It is easy to see that lines (1) and (2) do not have the same slope; line (2) is steeper than line (1). But how much steeper is it? How much more rapidly does it rise toward the 100.0 line? The equations expressing the slopes of the two lines placed beside each other show the difference in steepness immediately and precisely. The slope of a line is determined by the ratio

of the height of the ordinate to the length of the abscissa, and the formula is

$$m = y/x$$

What do these symbols mean? It is very simple. Any specified distance on OY (referring back to Figure XX) is designated as y. Any specified distance on OX is designated as x. In the figure y is the same as ON or MP, and x is the same as OM or NP. Now let us measure the length of these distances, y and x. The side of each small square will be assumed to be divided into 10 equal parts. They are found to be as follows:

y = 39.84%, or 39.84 small parts
x = 1,095 days, or 30 small parts

But it is the slope of the line in which we are interested. In order to find this, it is only necessary to substitute in the formula the values of x and y in terms of distance:

Hence, $m = 39.84/30.00$
$m = 1.328$, slope or tangent of angle MOP

The values of x and y must be expressed in terms of distance on the graph, and the slope, m, is found by dividing the value of y by the value of x.

Looking at the broken line and thinking of y as MP_1 and x as OM, we can compute the slope of the broken line in the same manner, as follows:

$m = y/x$
$m = 30.00/30.00$
$m = 1.000$, slope or tangent of angle MOP_1

Comparing the slopes of the two lines now, it is seen that the solid line is steeper by .328 than the broken line. The comparative steepness of two straight lines on different charts can be shown exactly by the formula above. An important fact about the slope of straight lines is that, if y is smaller than x, then m is less than 1. If they are the same size, then m is 1. If y is greater than x, as in this example, m is greater than 1.

Some straight lines cut the zero ordinate, OY, above the origin, O. How can the slope of a line which does that be computed? First, let us consider a problem in which this occurs. A sum of money is placed at simple interest, and the interest is allowed to accumulate. Table XII gives the accumulation of $1,000 at 6

per cent interest at the end of a 10-year period, and Figure XXI presents the data graphically (p. 149).

TABLE XII

THE ACCUMULATION OF $1,000 AT 6 PER CENT SIMPLE INTEREST AT THE END OF EACH YEAR OF A 10-YEAR PERIOD

Year	Principal Plus Interest
First	$1,060
Second	1,120
Third	1,180
Fourth	1,240
Fifth	1,300
Sixth	1,360
Seventh	1,420
Eighth	1,480
Ninth	1,540
Tenth	1,600

The line cuts the zero ordinate at M. Draw MN and NP, which in this problem are x and y respectively. The formula is the same as before:

$$
\begin{aligned}
m &= y/x \\
x &= 5 \\
y &= 1.2 \\
m &= 1.2/5 \\
m &= .24, \text{ the slope of MP}
\end{aligned}
$$

While this formula gives the slope of the line, it does not describe completely the line in its relations to the system of co-ordinates which is utilized in the construction of the graph. The general equation of a straight line, expressed in these terms, is

$$y = mx + b.$$

In this formula y equals the distance of any point on the line from the base line, or zero abscissa; m is the slope of the line; x is the length of the abscissa; and b is the distance between O and the point where the line cuts OY. The curve for the accumulation of a sum of money at simple interest is represented by an equation of the type of this formula. It is sometimes convenient to refer to a graph as of such and such a type, giving the equation instead of the graph.

The straight line graphs have many uses, most of which will be described later, because they enter into more complicated statistical methods. Regression lines in simple linear correlation are

straight lines and will be described in Chapter XI. Trends may be indicated by straight lines, and they will be described in Chapter XIII.

FIGURE XXI.—THE ACCUMULATION OF $1,000 AT 6 PER CENT INTEREST AT THE END OF EACH YEAR OF A 10-YEAR PERIOD

SEMI-LOGARITHMIC CHARTS

In the graphs which have preceded, the actual numbers have been plotted. But sometimes it is desirable to plot the logarithms

of the numbers instead of the actual numbers, at least on the vertical scale. A semi-logarithmic chart shows at a glance the *rate* of change, whereas the chart constructed from the actual numbers does not make this obvious. The semi-logarithmic chart has the natural numbers plotted on the horizontal scale and the logarithms of the second series plotted on the vertical scale. Ordinary graph paper may be used, in which case the worker looks up the logarithms for the series of numbers plotted on the vertical scale and shows only the logarithms for the series on the y-axis of the chart. Ratio, or semi-logarithmic, paper may be purchased which is ruled on the y-axis according to the logarithmic scale and obviates the necessity of looking up the logarithms.

The difference in appearance of data represented on the natural scale and on the logarithmic scale will be illustrated.

TABLE XIII

Population of the United States at Each Census, 1790 to 1930

Year	Population
1790	3,929,214
1800	5,308,483
1810	7,239,881
1820	9,638,453
1830	12,866,020
1840	17,069,453
1850	23,191,876
1860	31,443,321
1870	38,558,371
1880	50,155,783
1890	62,947,714
1900	75,994,575
1910	91,972,266
1920	105,710,620
1930	122,775,046

The data will be presented in two charts: the first uses the natural scale; the second uses the logarithmic scale along the y-axis.

Figure XXII shows the growth of population of the United States from 1790 to 1930 plotted on the natural scale. It indicates at a glance that the total population was small for the first five decades, but that after this point the aggregate number added every ten years increased markedly, and the largest single increase occurred from 1920 to 1930. The absolute increase has been higher for each succeeding decade except for the decades including

the Civil and World wars. But Figure XXII tells nothing about the *rate* of increase in each decade. Does the population of the United States show a correspondingly increased rate of growth?

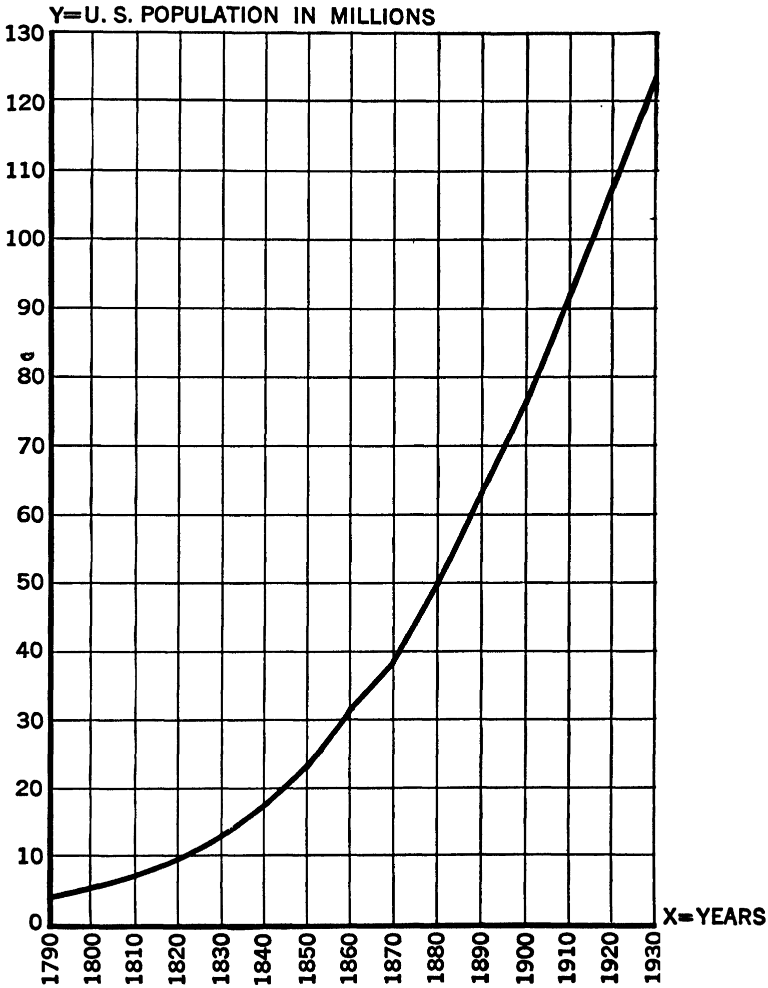


FIGURE XXII.—POPULATION OF THE UNITED STATES, 1790-1930 (NATURAL SCALE)

Figure XXIII, drawn to the logarithmic scale on the y-axis, answers this question. The rate of increase was larger in the earlier, instead of the later decades. The upper end of the curve shows a tendency to flatten out and to point to the time of an approxi-

mately stationary population in the not distant future. On the ratio chart the distance of the curve at any point from the base line

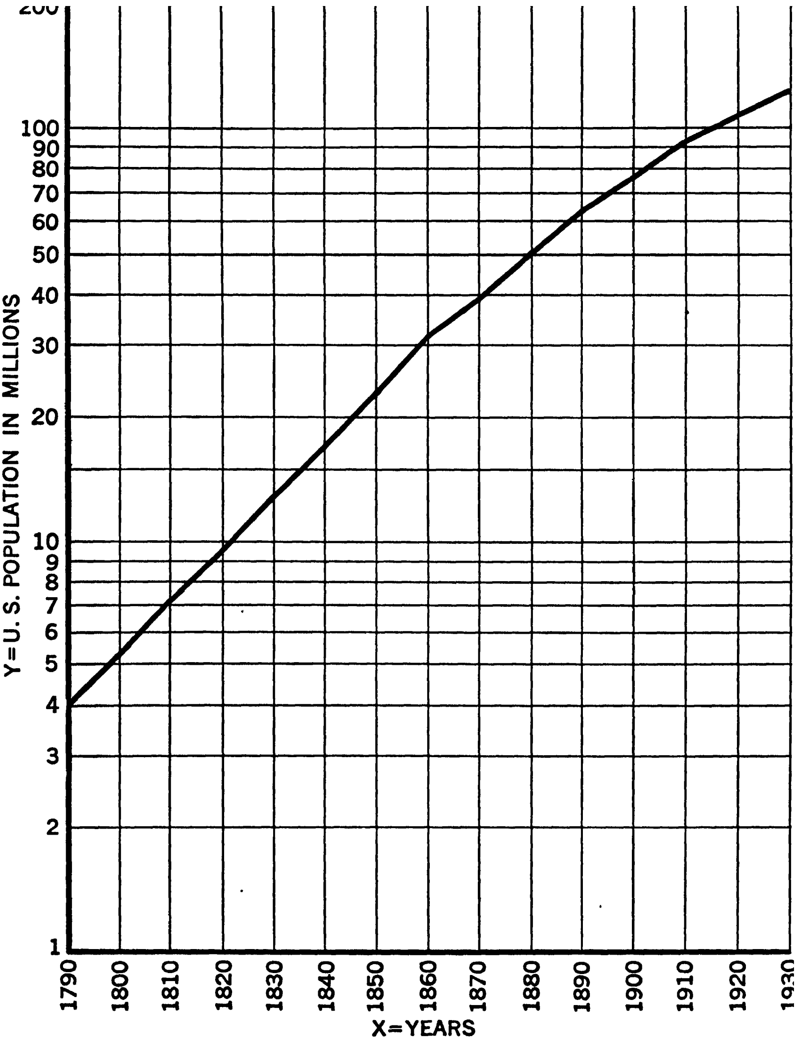


FIGURE XXIII.—POPULATION OF THE UNITED STATES, 1790-1930 (SEMI-LOGARITHMIC, OR RATIO, SCALE)

is not significant. Only the slope of the curve in Figure XXIII is significant; it indicates the rate of change. Frequently the rate of change in a series of social data is of primary importance, and the

aggregate increase may assume a secondary interest. In such cases the ratio chart should be used for presenting the facts.

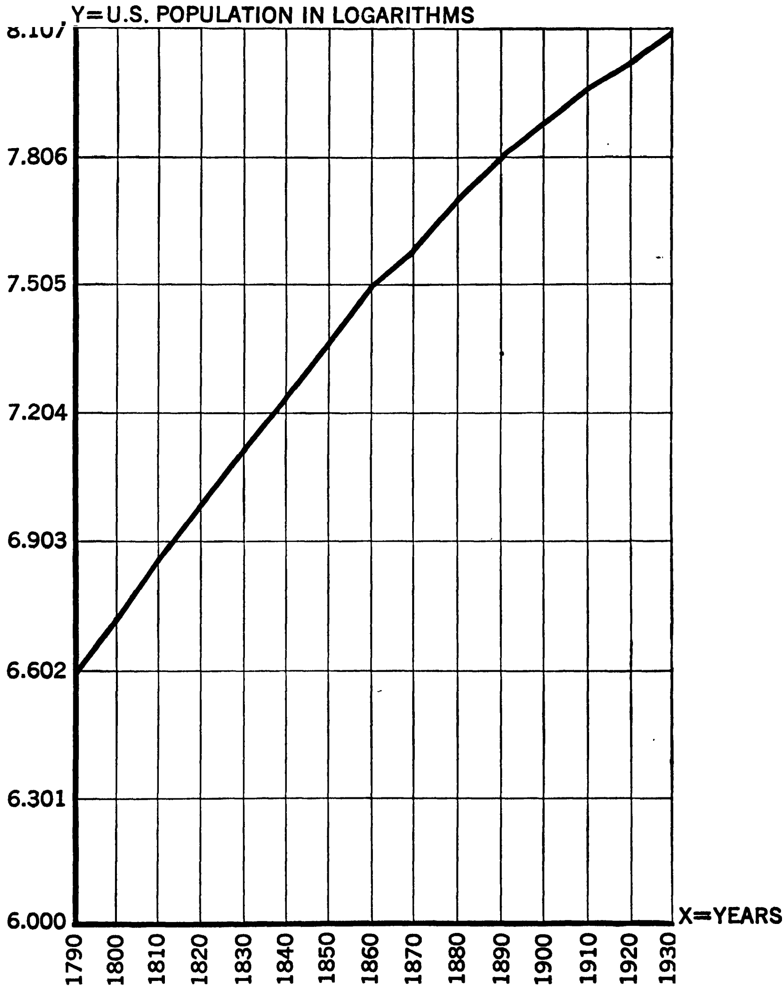


FIGURE XXIV.—POPULATION OF THE UNITED STATES, 1790-1930 (LOGARITHMS OF POPULATION PLOTTED ON THE VERTICAL SCALE)

Figure XXIV shows the same data on ordinary graph paper, but in this chart the logarithms of the numbers were found in a table and plotted, instead of the natural numbers.

The form of this curve is similar to that in Figure XXIII. However, the work can be done more rapidly, if ratio paper is used,

because that obviates the necessity of looking up the logarithms of the numbers.

In public welfare work administrators and the public are often interested in the rate of change from year to year. Figure XXV presents weighted index numbers of public welfare work in Indiana from 1900 to 1927 (p. 155).[2]

TABLE XIV

WEIGHTED INDEXES OF PUBLIC WELFARE WORK IN INDIANA, 1900 TO 1927

Year	Index	Year	Index
1900	88.9	1914	101.0
1901	88.1	1915	109.7
1902	88.7	1916	109.3
1903	89.2	1917	110.0
1904	92.9	1918	99.9
1905	93.8	1919	97.8
1906	94.9	1920	94.7
1907	92.6	1921	98.9
1908	95.5	1922	102.7
1909	92.3	1923	101.6
1910	95.3	1924	106.7
1911	97.0	1925	115.3
1912	99.2	1926	118.0
1913	100.0	1927	121.9

These index numbers vary from 88.1 in 1901 to 121.9 in 1927 which represents a large increase in the volume of work done (allowance was made in the index for increasing population and for changes in the purchasing power of the dollar), but the percentage change from one year to another is small. The volume of work is steadily growing, but the rate of increase is not large. The average increase was found to be a little less than 1 per cent a year, when the straight line trend was computed. The answer to the question of whether the ordinary chart or the ratio chart should be used depends upon the purpose of the worker. That should be clear before the form of presentation is decided.

6. CUMULATIVE CHARTS

In social planning it is necessary to estimate the probable volume of work and the necessary budget for 12 months or more in advance. In the case of budgets made on a biennial basis, such as those requiring appropriations from state legislatures or from

[2] These data are taken from "Indexes of Public Welfare in Indiana," by R. Clyde White, *Social Forces*, Vol. VIII, No. 2, p. 251.

Congress, it is necessary to plan two years in advance. If the need for the service has a general trend upward, then allowance must be made for a probably larger outlay in the second year than in the first year of the biennium. City departments also have to estimate their needs in advance. Once funds are available, the social

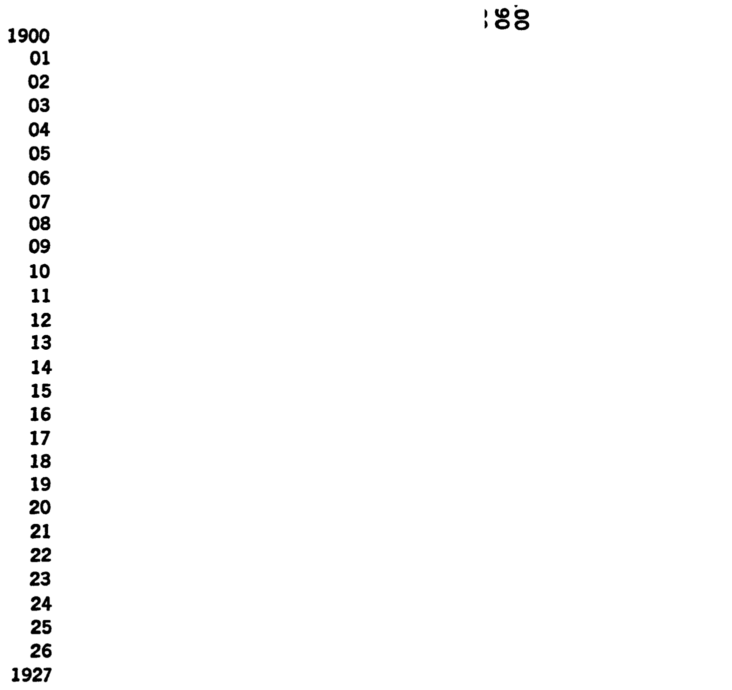


FIGURE XXV.—WEIGHTED INDEX OF PUBLIC WELFARE WORK IN INDIANA, 1900-1927 (SEMI-LOGARITHMIC SCALE)

agency or public department has to allot them on a monthly basis so that they will be distributed according to expected requirements. One of the statistical devices for keeping a close check on actual expenditures in relation to budgetary estimates is the cumulative chart.

An example drawn from the field of family case work will illustrate the value of the cumulative chart for such purposes. The expenditures of the Indianapolis Family Welfare Society were

obtained by months for a period of four years.[8] The average amount for each month was found for the four-year period, and then the percentage distribution by months was obtained. These percentages were cumulated by months and are represented in Figure XXVI by the solid line. The broken line shows the cumulated percentages through August, 1928, on the basis of an estimated relief budget of $50,000 for the year.

TABLE XV

CUMULATED PERCENTAGES OF ACTUAL EXPENDITURES BY MONTHS FOR 1928 AND CUMULATED PERCENTAGES OF BUDGET ESTIMATES FOR THE ENTIRE YEAR

Month	Percentages Cumulated, 1928	Percentages Cumulated, Estimates
All	86.3	100
January	13.1	12.1
February	27.3	24.1
March	42.0	35.4
April	52.2	44.1
May	61.4	51.5
June	69.8	57.9
July	77.8	64.3
August	86.3	70.6
September	....	76.4
October	....	82.6
November	....	89.9
December	....	100.0

This chart shows that the actual expenditures through August, 1928, were running steadily ahead of the budgetary estimate and that the funds would be exhausted before the end of the year. Such a situation is not uncommon in the history of relief agencies, because economic conditions cannot be predicted a year in advance. All the agency can do is to make as careful an estimate as possible and then make readjustments as new conditions are discovered. In the above case, either expenditures must be sharply reduced or additional funds obtained. At the end of any month in the year the relief agency could quickly see from the chart the financial problem it is facing. For presenting such data to boards of directors, the cumulative chart is very effective, and it is a useful guide to the executive who is trying to control expenditures by a monthly quota system. It will be obvious that the same kind of chart can be used to advantage by a manufacturer who plans his production

[8] Data from monthly reports of relief published currently by the Russell Sage Foundation, Department of Statistics.

for the year on a monthly basis, as a means of closely following the seasonal variations in the demand for his product and of

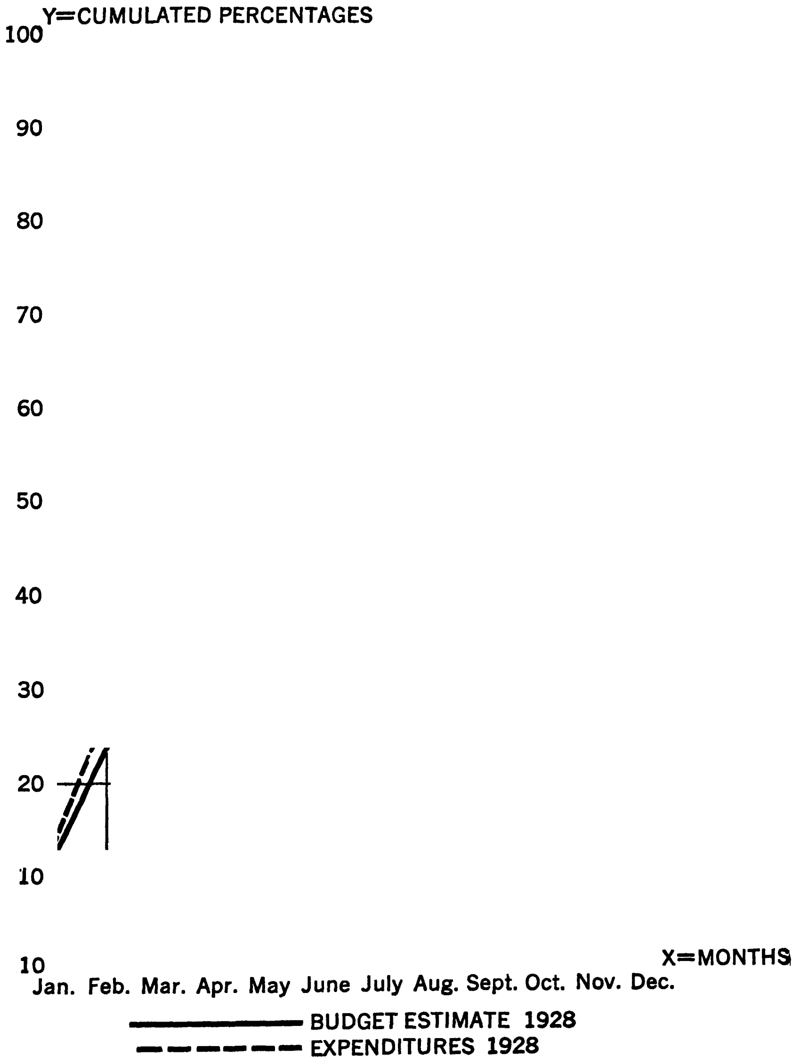


FIGURE XXVI.—COMPARISON OF BUDGETARY ESTIMATE AND ACTUAL EXPENDITURES IN 1928 THROUGH AUGUST, INDIANAPOLIS FAMILY WELFARE SOCIETY, IN TERMS OF CUMULATED PERCENTAGES

stabilizing production and employment. The cumulative chart acts as a sort of budgetary, or production, barometer.

This chart has still other forms and uses. Figure XXVII is constructed according to both the "more than" and the "less than" methods. These two methods can be explained best by reference to Table XVI:

TABLE XVI

FELONS SENTENCED IN THE MARION COUNTY CRIMINAL COURT, 1930, ACCORDING TO THE PERCENTAGE ABOVE (MORE THAN) OR BELOW (LESS THAN) A SPECIFIED AGE. 651 FELONS

Age	Per Cent of Felons More Than Specified Age	Per Cent of Felons Less Than Specified Age
16-	100.0	0.0
20-	70.2	29.8
25-	42.5	57.5
30-	31.1	68.9
35-	17.7	82.3
40-	10.2	89.8
45-	5.6	94.4
50-	2.4	97.6
55-	1.8	98.2
60-	1.0	99.0
65-	.5	99.5
70-	.2	99.8
75-	.0	100.0

One hundred per cent of all felons are "more than" 16 years of age, and none are "less than" 16 years of age. That is, all have passed the sixteenth birthday. Forty-two and five-tenths per cent have passed the twenty-fifth birthday, and 57.5 per cent have not reached the twenty-fifth birthday. If these two columns of percentages are plotted, they appear as in Figure XXVII (p. 159). To read the "more than" curve, look at any age on the horizontal scale, note the point on the ordinate erected from this point at which the "more than" curve cuts it, and read the percentage on the vertical scale opposite this point. This percentage is the percentage of felons at or above this age. The "less than" curve is read in a similar manner except that the percentage read is the percentage of felons who are less than this age.

In looking at this chart, it should be noticed that from 16 to 20 is a period of only four years, whereas from 20 to 25 is a five-year period. Hence, allowance is made for this fact in marking off the horizontal scale, and the distance from 16 to 20 is only four-fifths of the distance between the other figures.

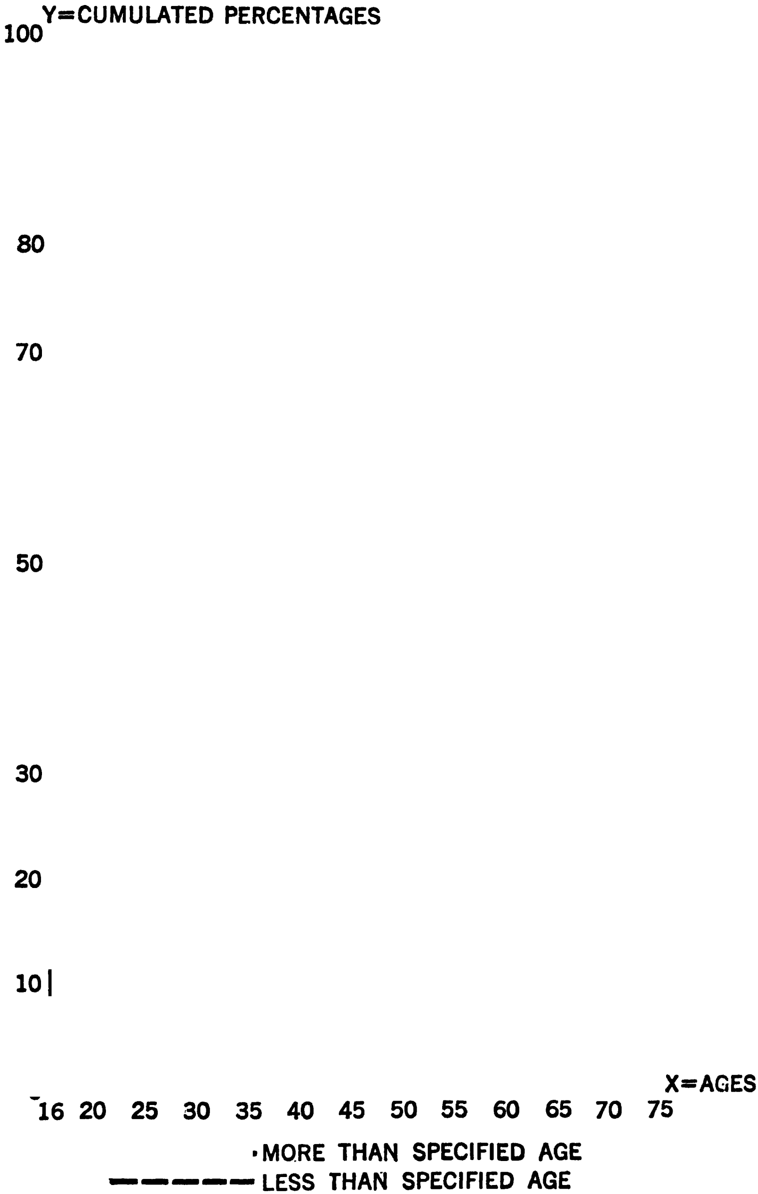


FIGURE XXVII.—CUMULATIVE CURVES SHOWING THE AGE DISTRIBUTION OF FELONS IN INDIANAPOLIS IN 1930 ON A "MORE THAN" AND ON A "LESS THAN" BASIS—651 FELONS

7. THE HISTOGRAM AND THE FREQUENCY POLYGON

The frequency distribution was discussed in Chapter VI, but it was presented there only as it appears in tables. Frequency distributions may be presented in graphic form also. Indeed, their graphic presentation is as common in statistical studies as is the tabular form. It was seen in Chapter VI that the array makes it possible for the statistician and the reader to gain a better idea of the meaning of a body of data than can be gained from the examination of an unorganized group of items. The array indicates the range of values from the lowest to the highest item. After discussing the array, it was shown how a still clearer idea could be obtained by grouping the data in class-intervals. This step prepared the material for presentation in the form of a frequency distribution. The histogram and the frequency polygon are two additional ways of reducing data to intelligible form. Mass data cannot be understood without resort to various devices for bringing out their meaning. The histogram will be considered first.

The first problem chosen to illustrate the use of the histogram is that of determining the age distribution of male employees in six moderately large firms: one department store, one street railway, and four factories.[4] Table XVII gives the data for the six firms in 10-year class-intervals:

TABLE XVII

Age Distribution of Male Employees in 6 Indianapolis Firms

Age	Number of Employees
All Ages	5,319
15–24	1,307
25–34	1,757
35–44	1,245
45–54	688
55–64	322

The age group with the greatest frequency is 25-34 years. This fact is obvious from the table, but it is made more emphatic by the chart on next page.

Each column in the histogram represents the number of employees at each age period. The concentration in the period, 25 to

[4] Data from an unpublished study by the author.

34 years, is marked, and the small number in the period, 55 to 64 years, is no less marked by the small height of the column.

The mechanics of the chart require some explanation. The columns have the same width, because each represents in the hori-

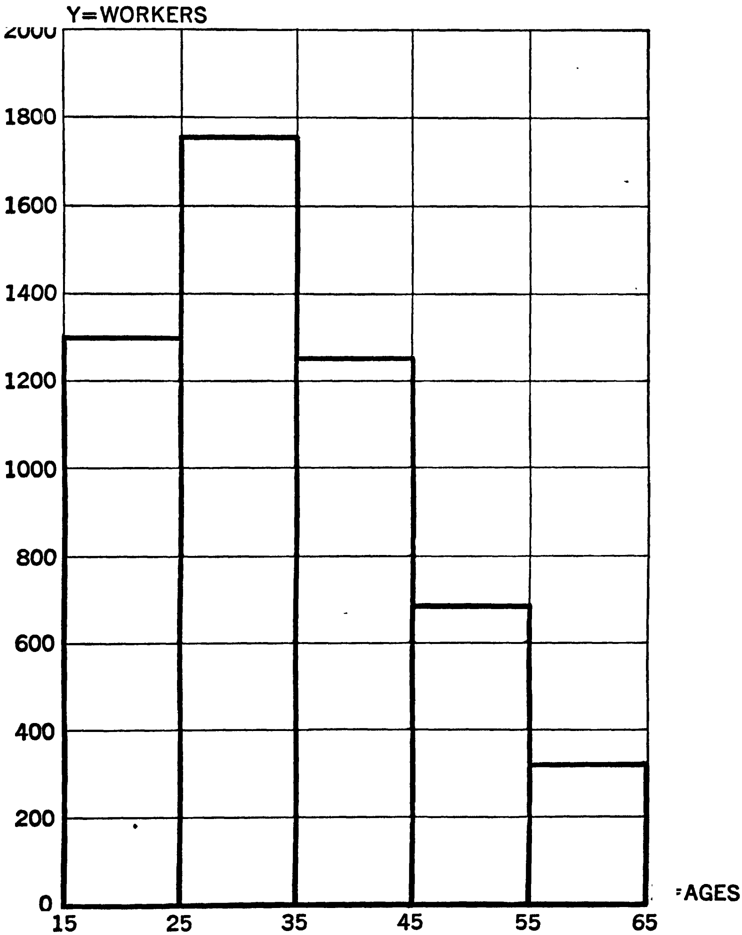


FIGURE XXVIII.—AGE DISTRIBUTION OF 5,319 WORKERS

zontal direction the same period of time. It is customary to plot the frequencies on the ordinates, that is, the vertical direction, and to chart the other variable, years in this case, on the abscissas, that is the horizontal direction. The table shows that in the first age group there were 1,307 men. At the termination of the distance on

the abscissa which is required for the first age period a vertical line is drawn upward until the end is opposite a point on the vertical scale equaling 1,307. From that point a line is drawn until it meets the zero ordinate at 1,307. It will be remembered from the discussion of class-intervals that the assumption is made that the items in the class-interval are distributed evenly from the lowest to the highest value. That assumption is made graphic here by the short horizontal line at the top of the column. It is parallel to the base line. If it were assumed that there are more items concentrated at the upper end of the class-interval than at the lower end, the line would slope upward toward the right. Later it will be shown that this is a fact in some distributions, but for such a rough presentation as the histogram provides it is unnecessary to give attention to this fact. The second column presents the number of men between 25 and 34 years of age. The right-hand side of the first column forms the left-hand side of the second column for a part of the distance, but it is prolonged to a point equal to 1,757 on the zero ordinate. At the upper terminus of this age group another vertical line is drawn equal in height to the first one. Then they are joined by a horizontal line. The other columns are formed in similar fashion. Thus, we have a comparison of the numbers of men in each age group, and we note the concentration in the second age group.

But does such a large class-interval make the meaning of the

TABLE XVIII

Age Distribution of Male Employees in 6 Indianapolis Firms and of the Total Male Population of Indianapolis for the Same Age Periods (Census of 1920)

Age Group	Male Employees		Males in Population	
	Number	Per Cent	Number	Per Cent
Total	5,386	100.0	114,111	100.0
15–19	263	4.9	11,516	10.1
20–24	1,044	19.4	14,936	13.1
25–29	1,002	18.6	15,675	13.7
30–34	755	14.0	14,361	12.6
35–39	715	13.3	14,000	12.3
40–44	530	9.8	10,385	9.1
45–49	371	6.9	10,426	9.1
50–54	317	5.9	8,824	7.7
55–59	202	3.8	6,051	5.3
60–64	120	2.2	4,752	4.2
65–69	67	1.2	3,185	2.8

data as clear as a smaller class-interval? In order to compare the visual effects of a chart which uses a smaller class-interval, the data have been divided into class-intervals of five years each. They

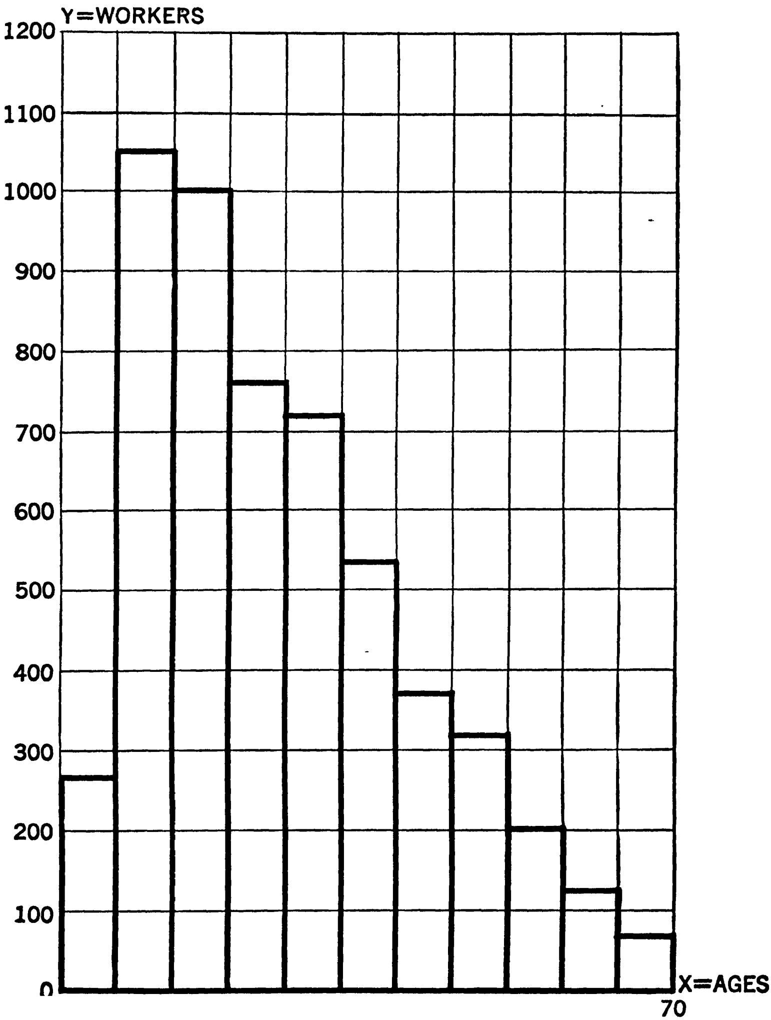


FIGURE XXIX.—AGE DISTRIBUTION OF WORKERS, 5-YEAR CLASS-INTERVALS

are presented in Table XVIII and are shown in Figure XXIX. The mechanics of construction are, of course, the same as in Figure XXVIII, but the impression one gets from the second chart is

that the number of employees in age groups above 20 to 24 years taper off more slowly than would be suspected from the first chart. In other words, while the first chart is technically correct and presents nothing but the facts, the large class-interval obscures the gradual decline in numbers in the higher age groups. This fact raises the question whether or not the male population in Indianapolis is distributed in a manner similar to that of the employees in Figure XXIX. The decline in numbers of employees is very gradual—almost in a straight line. Would a histogram of the males in Indianapolis between 15 and 64 years of age have a similar form? This raises a question which does not permit one to say that the employers are discriminating against the older man but suggests a further inquiry to clarify this point. Figure XXX presents two histograms together (data from Table XVIII): the solid line is a reproduction of Figure XXIX, and the broken line presents the age distribution of the male population in the entire city. The vertical scale of this chart is in terms of percentage instead of the actual numbers, because the population class frequencies are so much greater than those of the employees that an accurate comparison could not otherwise be made between the two series.

Some significant differences appear between the two histograms of Figure XXX. The number of men in the upper age groups does gradually decrease, but the largest percentage in any age group does not reach the largest percentage in certain age groups of the employees. After the forty-fifth year the percentage of men in the population in each age group exceeds the percentage employed by the six firms. It is clear, then, that younger men predominate in these firms and that either older men are not accepted as readily as younger men or they do not attract the older men. The last possibility suggests that still further inquiry is necessary, before it is possible to decide whether these firms discriminate against older men. As a matter of fact, the older men are found in relatively smaller numbers. Would they be employed, if they applied for jobs? The data presented here are inadequate to answer that question. The histograms merely present the conditions as they exist, which is their purpose.

The preceding histograms have shown that the high percentages of employees in the six firms come at the earlier working ages. Some social data are distributed in different ways. It is common to find social series which have the concentration at the middle and

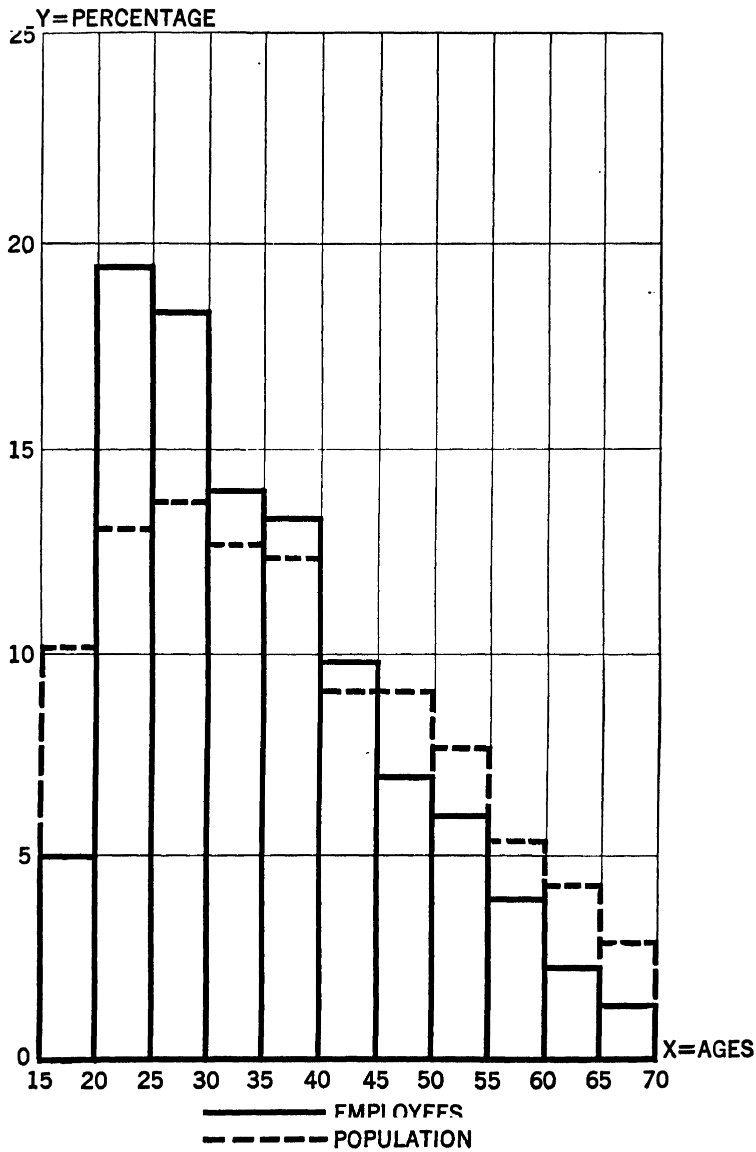


FIGURE XXX.—COMPARISON OF THE AGE DISTRIBUTION OF EMPLOYEES IN SIX FIRMS AND OF THE TOTAL MALE POPULATION OF INDIANAPOLIS BETWEEN 15 AND 64 YEARS OF AGE IN TERMS OF PERCENTAGE

also at the upper end of the scale. Death rates by age groups show concentration at both ends of the age scale. An illustration will be given of data which concentrate near the center of the scale. Table XIX gives the distribution by ages of children in the eighth grade in the St. Louis Public Schools.

TABLE XIX

DISTRIBUTION OF CHILDREN IN THE EIGHTH GRADE, ST. LOUIS PUBLIC SCHOOLS, BY AGES[1]

Age	Number of Children
All Ages	4,721
10	1
11	25
12	348
13	1,330
14	1,684
15	971
16	308
17	50
18	4

[1] Data from Woodrow, Herbert, *Brightness and Dullness in Children*, Lippincott, Philadelphia, 1919, p. 130.

In Figure XXXI the data are presented in the form of a histogram to show the concentration at the middle of the age scale (p. 167). This histogram is almost symmetrical: it rises steeply from the lowest age group to the middle age group, and then declines rapidly to the highest age group. Fourteen is the most common age of children in the eighth grade. Comparatively, those in the lower age groups are advanced, and those in the higher age groups are retarded. It should be noticed that considerably more children are advanced by one year than are retarded. In view of the fact that the children who are two, three, and four years advanced or retarded are about equal in each age period, it may be that there is some artificial factor, such as an administrative practice, operating to make the unevenness in numbers of those who are advanced or retarded only one year. The fact that the histogram is symmetrical but for the differences at these two ages suggests that the numbers advanced or retarded only one year would be more nearly equal, if their status depended upon their ability alone, or possibly even if we had an indefinitely large number of children for this grade to study.

In connection with histograms and frequency curves the consideration of discrete and continuous variables is apropos. It will be

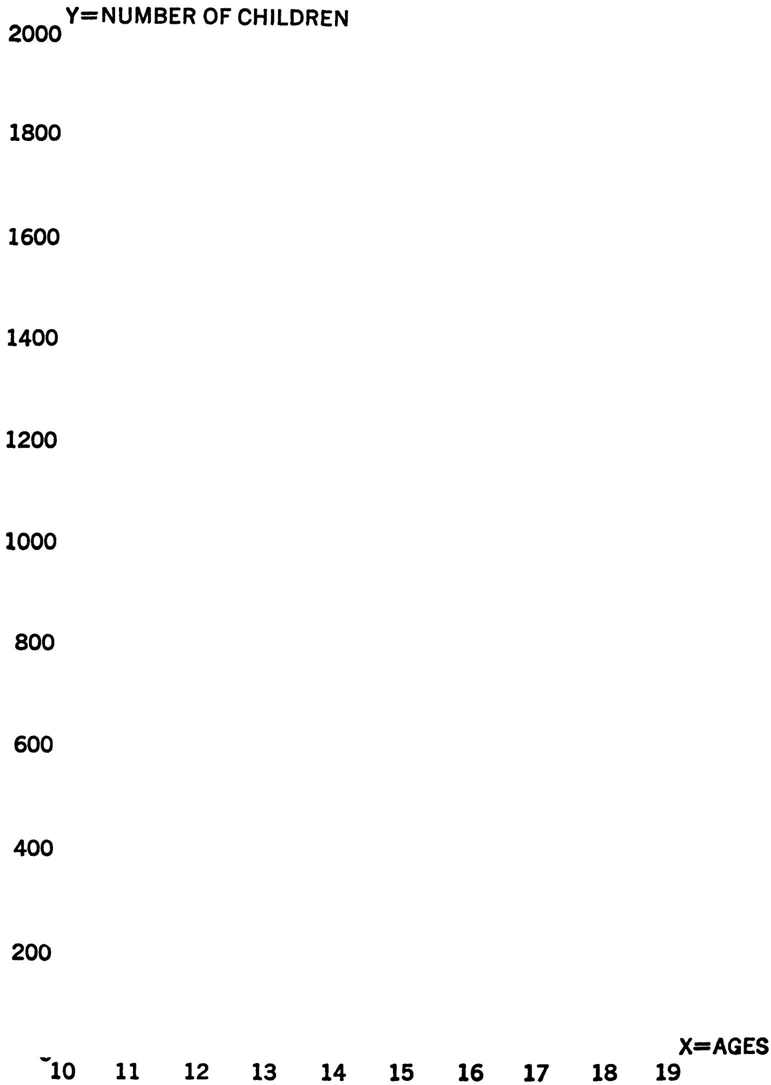


FIGURE XXXI.—DISTRIBUTION OF CHILDREN IN THE EIGHTH GRADE, ST. LOUIS PUBLIC SCHOOLS, BY AGES

recalled from Chapter III that a discrete variable was defined as one whose values differed by an assigned amount, whereas the

values of a continuous variable differed by infinitely small amounts. Logically the discrete variable ought to be presented graphically only by a histogram, because a histogram is a form of column chart and does not suggest continuity in the series of data. The values of a discrete variable show gaps, sometimes small, in the series arranged as a frequency distribution. In practice these often approach the form of an ideal frequency curve and are so presented. If this is done with complete understanding that the variable is really discrete and if no attempt is made to draw from it inferences which could apply only to a continuous series, the practice is not objectionable. The continuous series may properly be presented as a smooth curve, because, even if the data do show small gaps, the variable changes by amounts infinitely small and, if a sufficient number of items were included, would take the form of a smooth curve. The histogram may be used to present continuous variables, as in Figure XXXI above, where the independent variable, age, may vary by any amount however small, but it may err on the side of suggesting that the variable is discrete when it is not. This practice is also permissible, if it is clear to the worker that his data really should be presented as a frequency curve and if no misunderstanding would result.

The data in Table XIX are a continuous series and may be used to illustrate the development of a smooth frequency curve from a histogram. As pointed out, this distribution is approximately symmetrical in form and approaches the form of distribution represented by a "normal" or a bell-shaped curve. In scientific work this type of curve has many uses, and when we come to discuss the theory of probability in Chapter XII, it will receive extended attention. Intermediate between the histogram and the smooth frequency curve is the *frequency polygon.* The histogram is composed of a number of vertical bars. If the mid-points of the tops of these bars are connected by straight lines, the result is a frequency polygon. Figure XXXII illustrates this point.

A polygon is commonly defined as a geometrical figure having more than four angles. The angles of the above polygon are located at the mid-points of the tops of the vertical bars and are formed by the straight lines which connect the mid-points. It will be noticed that the form of the frequency polygon emphasizes better than the histogram the concentration of children at the middle age period and also the approximately symmetrical dis-

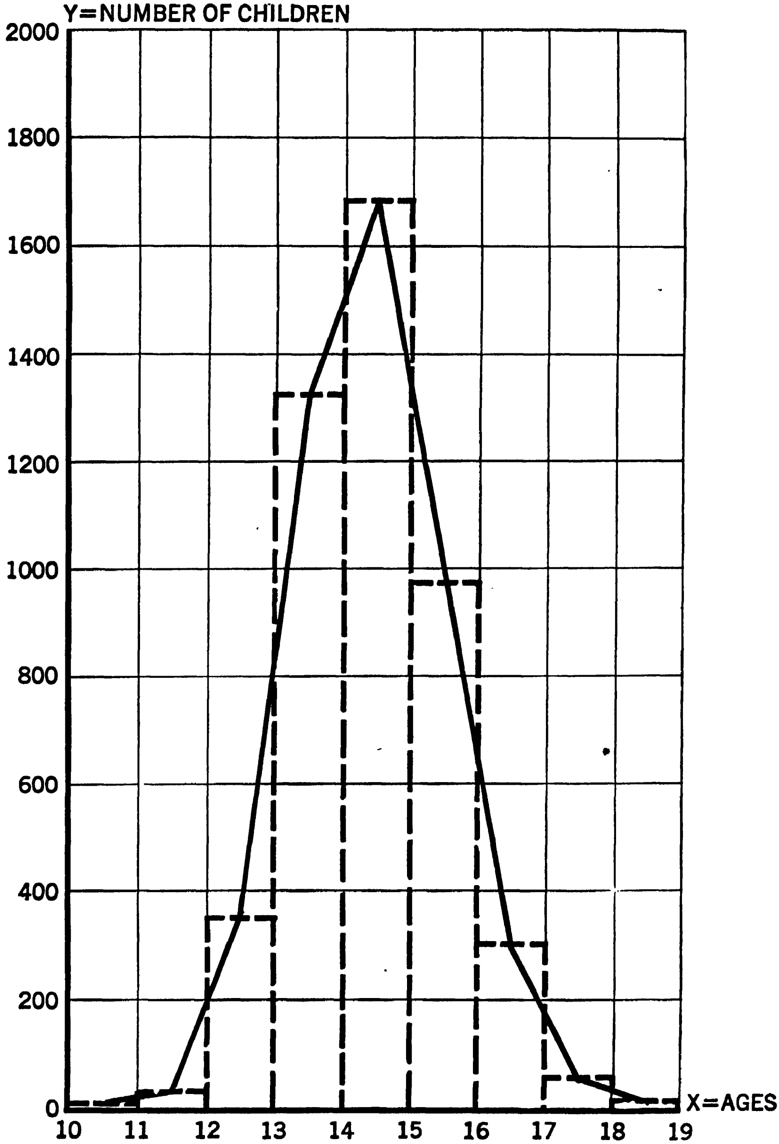


FIGURE XXXII.—DISTRIBUTION OF CHILDREN IN THE EIGHTH GRADE, ST. LOUIS PUBLIC SCHOOLS, BY AGES, SHOWING THE RELATIONS BETWEEN A HISTOGRAM AND A FREQUENCY POLYGON

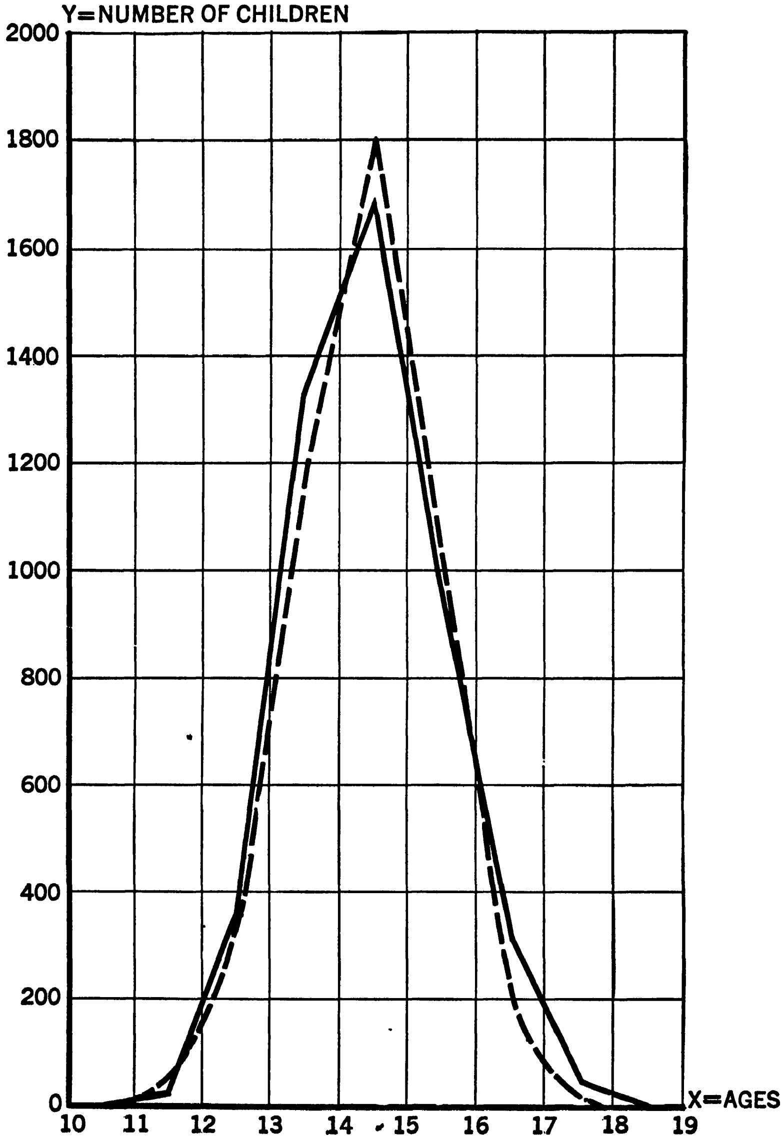


FIGURE XXXIII.—DISTRIBUTION OF CHILDREN IN THE EIGHTH GRADE, ST. LOUIS PUBLIC SCHOOLS, BY AGES, COMPARING THE FREQUENCY POLYGON AND THE SMOOTHED FREQUENCY CURVE

tribution of children below and above the point of concentration.[5] If the purpose is further to emphasize the symmetry of the distribution, the frequency polygon may be smoothed by drawing a free-hand line around the polygon. This is illustrated in Figure XXXIII.

This smoothed frequency polygon is still not entirely symmetrical. It bulges a little on the left side, and it is pushed inward a little on the right side. There are two possible explanations of why these data distribute themselves in this slightly asymmetrical way: this may be the normal distribution of children of different ages in the eighth grade, or an artificial factor may be producing the asymmetry. A much larger number of children might tend to remove the apparent asymmetry. Before drawing any conclusion regarding the natural distribution of such data, that is, before one can determine the statistical law describing them, further experimentation is necessary. Nevertheless, it is interesting to place a symmetrical curve on the polygon in Figure XXXII to see in what respects it differs from the actual distribution. This symmetrical curve, when smoothed by proper methods, is known as an ideal frequency curve. Chapter XII, which deals with the theory of probability, will describe methods for fitting an ideal curve to any frequency distribution of this general type. For the present it will suffice to see how the two curves look when superimposed in Figure XXXIV.

If the children were evenly distributed on both sides of the arithmetic average, the distribution would be entirely symmetrical and would be represented by the broken curve in Figure XXXIV. In the three preceding charts it will be apparent that there is a gradation from the histogram to the ideal frequency curve. The histogram is the simplest representation of the data. The frequency polygon is still close to the original data, but the free-hand smoothed frequency curve is a step further away from the data. The ideal frequency curve is entirely theoretical; it represents the form which the distribution of the 4,721 children would take if they conformed to the ideal distribution. It is useful for comparing the departure of the actual data from the ideal distribution, or, as it is sometimes called, the "normal probability" curve. If a

[5] In this as in other line graphs, the left end of the base line along which the horizontal scale is measured off is referred to as the lower end of the scale, and the right end of the base line is correspondingly referred to as the upper end of the scale. The terms "below" and "above" the point of concentration in the frequency distribution are used in the same manner.

larger number of children, in this case, were used and their distribution approached nearer and nearer the ideal frequency distribution, it might be correct to say that the ideal curve is the general

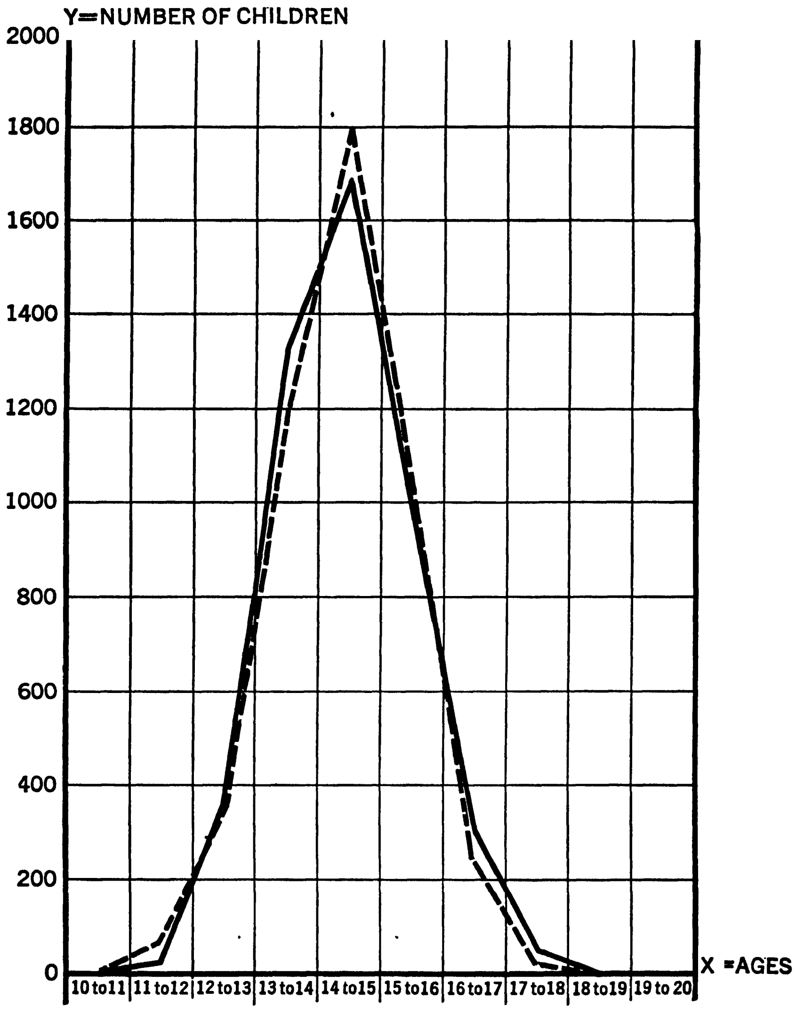


FIGURE XXXIV.—DISTRIBUTION OF CHILDREN IN THE EIGHTH GRADE, COMPARING THE FREQUENCY POLYGON WITH THE IDEAL FREQUENCY CURVE

statistical law which describes the data. Then any particular study of the age distribution of eighth grade children which failed to approach the ideal curve would obviously be an unrepresentative study, either because not enough children had been included or

because some other factor had entered into the situation to skew the curve.

The next frequency chart shows how the component parts of a population problem may be represented graphically. This chart is taken from a study of school attendance covering the entire United States. Do all of the major population groups analyzed according to nativity and race show the same percentage of children in school at different ages? When the data were assembled and analyzed,

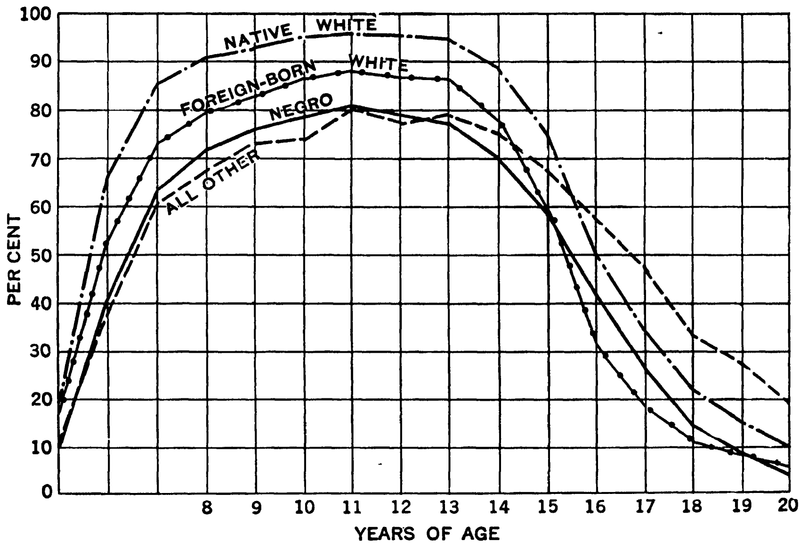


FIGURE XXXV.—PER CENT OF MALES ATTENDING SCHOOL AMONG THE NATIVE WHITE, FOREIGN-BORN WHITE, NEGRO, AND "ALL OTHER" POPULATION 5 TO 20 YEARS OF AGE, BY SPECIFIED AGE: 1920[5a]

differences were found. Figure XXXV shows the situation. The native white population shows the highest school attendance until after the fifteenth birthday. From that point on the "all other" ranks highest. The Negro group is generally low, but after the fifteenth birthday the foreign-born white children drop below all the others. The chart makes possible comparisons at any age between any two or more population groups, and as a whole gives a general impression of how these groups attend school.

It will be found in practice that only a small proportion of

[5a] Ross, Frank A., *School Attendance in the United States, 1920,* p. 8. United States Bureau of the Census, 1924.

series of social data are distributed in the approximate form of the ideal frequency curve. Most frequency distributions in the field of social statistics will lack the perfect symmetry exhibited by the bell-shaped curve. They will not be nearly so symmetrical as the frequency polygon constructed from data for the age distribution of eighth grade children. In fact, they will be noticeably asymmetrical; that is, they will look as if they had been pushed toward one side or the other. Using normal in the sense of average, these asymmetrical distributions may be normal for the data used. It may be possible to fit a smoothed, or generalized, curve to these asymmetrical distributions to which all sample studies would closely conform. Table XX gives the number of cities in the United States which had increased less than 120 per cent between 1920, and the time of the 1930 census, distributed in 10 per cent class-intervals:

TABLE XX

336 Cities in the United States, with 25,000 or More Population, Which Increased Less Than 120 Per Cent between 1920 and 1930

Percentage Increase	Number of Cities
Total	336
Under 10	69
10– 19	78
20– 29	62
30– 39	41
40– 49	23
50– 59	20
60– 69	7
70– 79	16
80– 89	4
90– 99	6
100–109	6
110–119	4

The most common rate of increase lies between 10 and 20 per cent. Few cities showed an increase of over 40 per cent. It should be noted that a few cities lost population, while a few had increases greater than 120 per cent. The number which had decreases is about equal to the number which had more than 120 per cent increases; so for convenience in constructing the chart these extremes were omitted. But they would have to be included if one were computing the average change in population of this class of cities. For the data presented the concentration is at the lower end of

FIGURE XXXVI.—336 CITIES IN THE UNITED STATES, WITH 25,000 OR MORE POPULATION, WHICH INCREASED LESS THAN 120 PER CENT BETWEEN 1920 AND 1930

the scale. For some other series, such as the age distribution of persons dying of heart disease, the concentration would be at the upper end of the scale.

Figure XXXVI has one technical difference from preceding charts. It shows percentage as the independent variable and, hence, plotted on the horizontal scale. In general percentage will be the dependent variable, but in this case the maximum percentage is arbitrarily fixed, and the class-intervals are fixed. But the number of cities in each class is not fixed; the only way it can be determined is to count the cities falling into each class frequency. This number varies according to the length of the class-interval, which here is 10. In this problem we are concerned with the frequency of cities in percentage groups; this fact determines which is the independent and which the dependent variable.

8. MISCELLANEOUS GRAPHIC DEVICES

Besides such curves as have been discussed, there are a great many devices used for graphic presentation. Some will be illus-

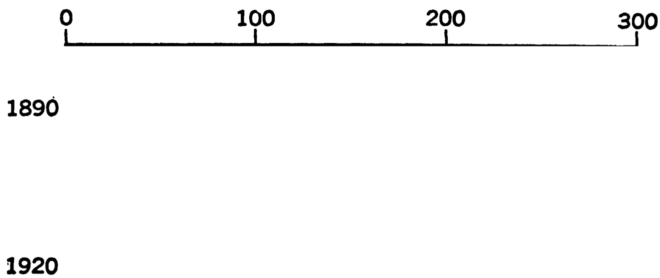


FIGURE XXXVII.—REPRESENTING THE PERCENTAGE OF CHANGE IN POPULATION OF INDIANAPOLIS FROM 105,436 IN 1890 TO 314,194 IN 1920

trated at this point, but before giving the illustrations a caution must be mentioned. It was pointed out earlier that graphic methods are ways of translating concrete data into symbols expressed as lines, surfaces, and sometimes cubes. Where equations are presented along with the graph, any of these geometrical concepts may be employed, though surfaces are more apt to be misleading than lines, and cubes more difficult to utilize with clarity than surfaces. It is easier for the eye to grasp the relative size of two lines of varying lengths than two surfaces of varying areas or two solids of varying cubic content.

For example: The city of Indianapolis increased in population from 105,436 in 1890 to 314,194 in 1920, almost exactly tripling the population. Let us represent the growth of the city, first, by narrow bars (a bar is a narrow space enclosed between two lines

17.3

10.0

1920
298%

1890
100%

FIGURE XXXVIII.—REPRESENTING PERCENTAGE CHANGE IN POPULATION OF INDIANAPOLIS FROM 105,436 IN 1890 TO 314,194 IN 1920 BY MEANS OF AREAS

the length of which is so obvious that width is neglected by the eye, and not used in fact); second, by squares; third, by cubes.

Although it might not be known what the exact population was in 1890 and 1920, a glance at Figure XXXVII would immedi-

6.7

4.6

1890
100%

1920
298%

FIGURE XXXIX.—REPRESENTING PERCENTAGE CHANGE IN POPULATION OF INDIANAPOLIS FROM 105,436 IN 1890 TO 314,194 IN 1920 BY MEANS OF CUBES

ately suggest that in the thirty-year period the population had just about tripled. But an examination of Figure XXXVIII, in which the area of the large square is three times that of the small square, does not suggest to the eye a tripling of the population; the second

square does not appear to be three times the size of the small one. When cubes are used in Figure XXXIX, the relative magnitudes of the two cubes are still less obvious. Sometimes a small man and a large man are used to illustrate growth in population, but this is a special case of the cubic representation, because the figure of a man is three dimensional, albeit irregular in contour. The illusion would be present, even though only the height of the men were intended for comparative purposes, because the other dimensions of the large man would give the effect of less height than he possessed in fact. Clearness will be enhanced in graphic presentation of this sort if comparisons are made by one dimension only. There may be special cases where the square, the rectangle, or the cube is most satisfactory, but they do not occur often.

The bar chart, or the column chart which differs from the bar chart only in the fact that the bars are erected vertically on the base line, is one of the simplest and most easily understood of all graphic devices. Instead of using only two years, as in Figure XXXVII, one may use the bar chart to compare the population at the time of the census in several different years. For non-technical and non-functional presentation of statistical data the bar chart is widely used. It does not take the place of more refined statistical analysis but is satisfactory for presenting the elementary implications of some data.

Variations of the bar chart are the hundred-per-cent chart, Figure XL, and the double-bar chart, Figure XLI, below:

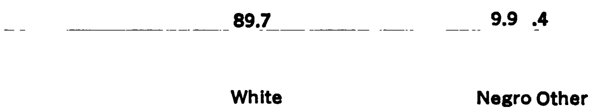


FIGURE XL.—PERCENTAGE OF THE POPULATION OF THE UNITED STATES REPRESENTED BY EACH RACE, 1920

TABLE XXI

PERCENTAGE OF THE POPULATION OF THE UNITED STATES REPRESENTED BY EACH RACE, 1920[1]

Race	Percentage of Population
White	89.7
Negro	9.9
Indian	.2
Chinese	.1
Japanese	.1

[1] *Abstract of the Census, 1920.*

The percentages of Indians, Chinese, and Japanese are so small that they were combined and represented as "other." The total length of the bar is 100 per cent. Hence, the division into parts representing the proportions of different races in the population shows the relative importance of the races.

Figure XLI illustrates the use of the double-bar graph. The data are taken from criminal statistics and are given in Table XXII:

TABLE XXII

PERCENTAGE OF WHITE AND NEGRO RACES AMONG THE COMMITMENTS TO PRISONS AND REFORMATORIES, 1910 AND 1923.[1]

Race	Commitments in Per Cent	
	1910	1923
All	99.2	98.2
White	66.3	74.2
Negro	32.9	24.0

[1] *Prisoners, 1923.* Report of the United States Bureau of the Census.

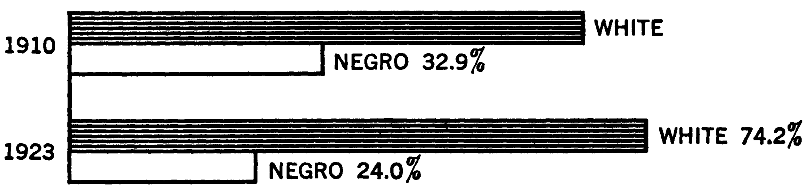


FIGURE XLI.—PERCENTAGE OF WHITE AND NEGRO RACES AMONG THE COMMITMENTS TO PRISONS AND REFORMATORIES, 1910 AND 1923

This graph presents the relative percentages of commitments of whites and Negroes to prisons and reformatories in 1910 and 1923 and brings out the point that Negroes have declined in proportion to whites in commitments to penal institutions of the United States.

A variation in bar charts is shown below. The purpose of this graph is to compare the percentage distribution by ages of the total population of the United States in 1920 and of the gainfully employed in the same year to emphasize the age-group variations. This kind of bar chart makes clear the relation of the occupied group to the total population. Few persons under 16 years of age are employed, but in the next two age groups the percentages gainfully employed are higher than the corresponding percentages

TABLE XXIII

AGE DISTRIBUTION OF THE POPULATION OVER 10 YEARS OF AGE AND OF THE GAINFULLY EMPLOYED OF SIMILAR AGES EXPRESSED IN PERCENTAGE [1]

Age	Percentage of Population	Percentage of Gainfully Employed
All	100.0	100.0
10–15	14.9	2.5
16–44	57.5	69.4
45–64	21.5	23.8
Over 65	5.9	4.1
Unknown	.2	.2

[1] *United States Census of Occupations, 1920.*

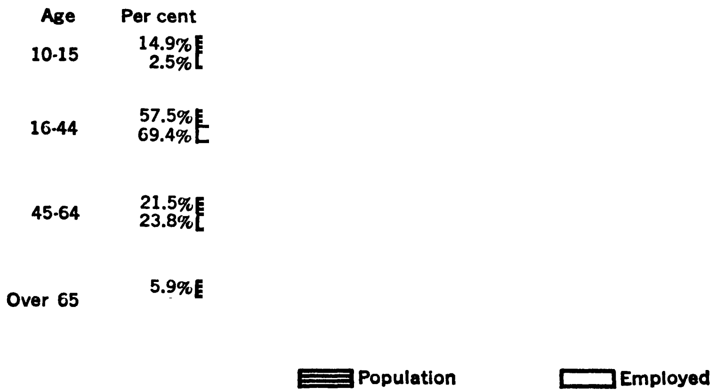


FIGURE XLII.—AGE DISTRIBUTION OF THE POPULATION AND OF THE GAINFULLY EMPLOYED OVER 10 YEARS OF AGE

of the total population above 10 years of age in these two age groups. In the upper age group the percentage employed falls below the corresponding percentage of the population. This kind of chart may be used to advantage in comparing the age distribution of persons who receive the services of social agencies with similar age groups in the population.

The circle, or sector, chart (otherwise called the pie chart) resembles both the surface chart and the hundred-per-cent chart in that the area bounded by an arc and two radii is used and that this area is a part of the whole circle which is 100 per cent. This kind of surface chart does not have the disadvantages of the quadrilateral or the triangle, because the whole circle is conceived as representing all the data, and the sectors are parts of this whole.

This relationship brings out the relative magnitude of each division. Figure XLIII illustrates this type of graph:

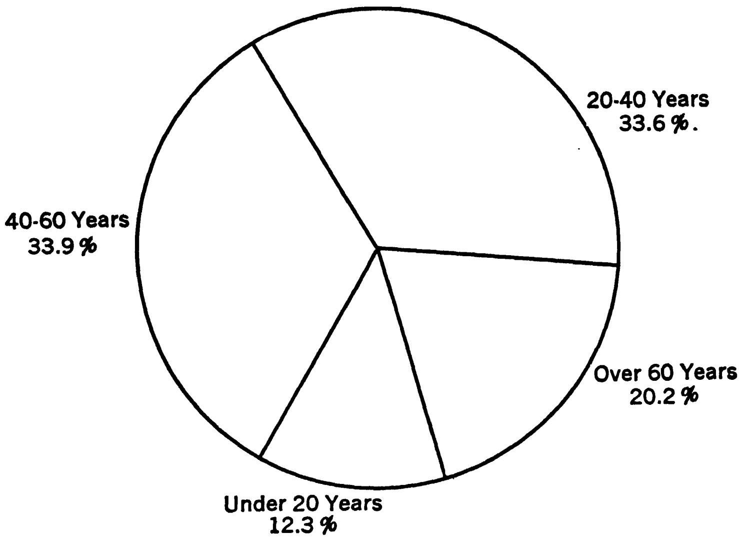


FIGURE XLIII.—NEW COMMITMENTS TO INDIANA HOSPITALS FOR THE INSANE BY AGE GROUPS, YEAR ENDING SEPTEMBER 30, 1929

TABLE XXIV

NEW COMMITMENTS TO INDIANA HOSPITALS FOR THE INSANE BY AGE GROUPS, YEAR ENDING SEPTEMBER 30, 1929[1]

Age	New Commitments	Per Cent
All	1,642	100.0
Under 20	201	12.3
20–40	552	33.6
40–60	557	33.9
Over 65 and Unknown	332	20.2

[1] *Indiana Bulletin of Charities and Corrections*, No. 182, p. 185.

The division of the circle into parts showing each age group's proportion of the total commitments makes clear the age groups from which the hospitals for the insane draw most of their patients. It is probably a better form than the hundred-per-cent bar, and it does not have the objectionable features characteristic of rectangular surfaces.

Geographic data are often satisfactorily presented by the use of

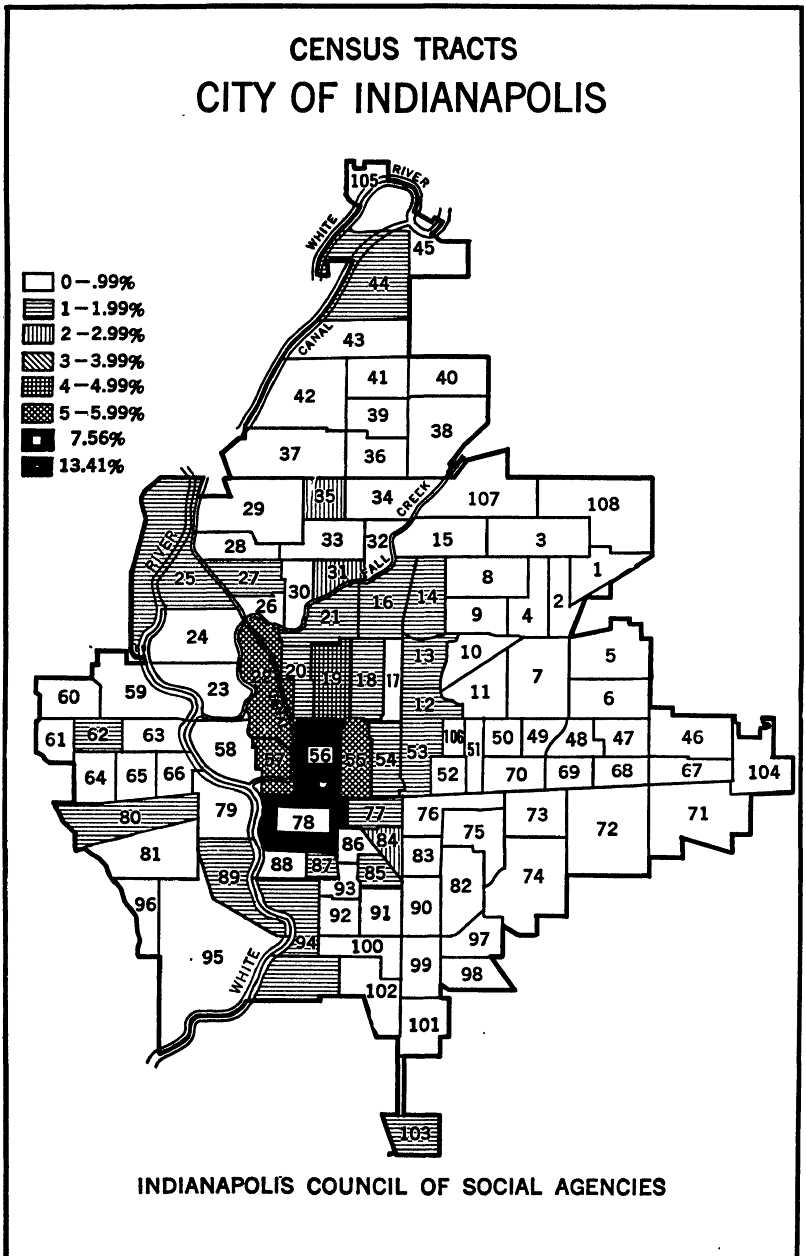


FIGURE XLIV.—LOCATION OF FELONIES, JANUARY TO JUNE, 1929

a map of the area from which the data are taken; this is divided into small subdivisions, such as states for the United States, counties for a state, townships for a county, or wards or census tracts for cities. Such maps are used in the so-called ecological studies of social problems which have been made in Chicago and elsewhere. Since the Bureau of the Census began tabulating some of the population data for cities by "census tracts," the use of maps to present the distribution of disease, crime, poverty, etc., has greatly increased. The population of the tracts is small and usually highly homogeneous as to race, nationality, economic status, age, and sex. If the data of crime, disease, and poverty are distributed by census tracts, it is possible to make important studies of the occurrence of social-problem phenomena. To a lesser extent counties may be used to study problems on a state-wide basis. Figure XLIV is a good illustration of this use of the map. (See p. 182.) Figure XLIV shows the percentage of convicted felons whose crimes were committed in each census tract of Indianapolis from January to June, inclusive, 1929. Tracts 56 and 78 include the main business part of the city, and it will be noticed that 20.97 per cent of all felonies were committed in these two tracts. Some other contiguous tracts also had high rates of crime. It is clear that police protection should be concentrated in tracts 56 and 78 and to a lesser extent in a few other tracts. When these data are related to other facts obtained by the census, additional inferences of importance may be made.[6]

A variation of the cartogram is shown in the next chart. This is taken from a recreation study made in Indianapolis:

TABLE XXV

DISTRIBUTION OF HOMES OF CHILDREN USING A PUBLIC PLAYGROUND[1]

Distance of Home from Playground	Number of Homes	Percentage of Homes
All	146	100.00
Under ¼ mile	86	58.90
¼ to ½ mile	36	24.65
Over ½ mile	15	10.27
No address	9	6.18

[1] Lies, Eugene T., *The Leisure of a People*, Indianapolis Council of Social Agencies, 1930, p. 132.

[6] Data for this map were collected by the writer but have not been published.

The investigator wanted to determine the soundness of the present distribution of playgrounds in Indianapolis from the point of view of maximum use. For three days he adopted the plan of putting some one in the playground to tag every child and to learn his home address. Then each home was indicated by a dot on the map. When this was completed, he drew a circle with a

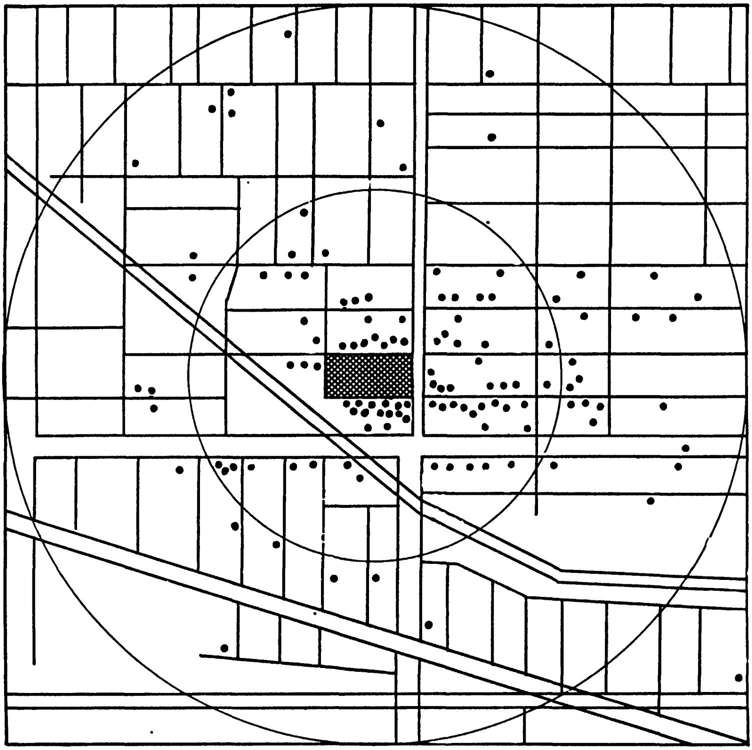
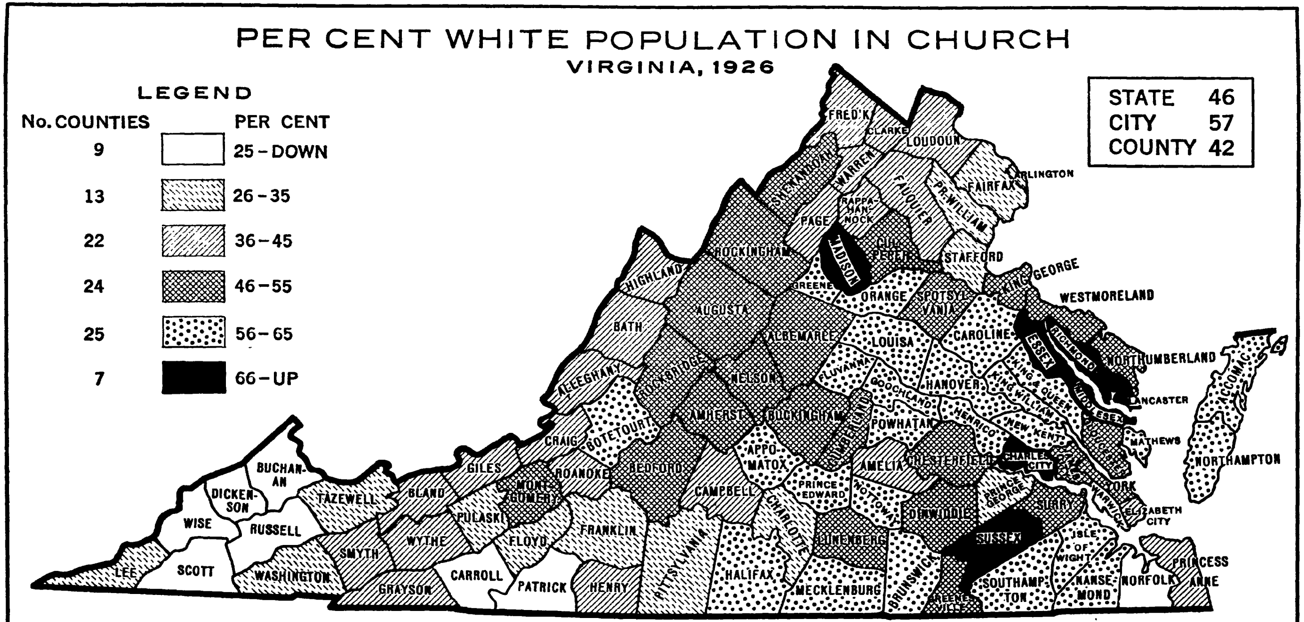


FIGURE XLV.—DISTRIBUTION OF HOMES OF CHILDREN USING A PUBLIC PLAYGROUND SHOWN BY ONE DOT FOR EACH HOME AND BY CONCENTRIC CIRCLES OF A QUARTER-MILE AND A HALF-MILE RADIUS

radius of a quarter of a mile from the center of the playground. Then a concentric circle with a half-mile radius was drawn. The homes inside the first circle were counted; then those lying outside the first but inside the second circle were counted. This method gave a basis for estimating how far children would go to reach a playground and where future playgrounds ought to be located. The type of cartogram used by Mr. Lies has great usefulness,



Staunton	82	Suffolk	73	Clifton Forge	66	Roanoke	57	Portsmouth	51
Lynchburg	79	Charlottesville	69	Petersburg	63	Norfolk	55	Buena Vista	37
Harrisonburg	79	Winchester	69	Richmond	60	Alexandria	51	Radford	26
Bristol	78	Fredericksburg	67	Danville	59	Newport News	51		

Source: U. S. Census of Religious Bodies, 1926. Population Estimated, 1926.

FIGURE XLVI.—PERCENTAGE OF THE WHITE POPULATION IN COUNTIES OF VIRGINIA WHO BELONG TO CHURCHES[1]

[1] Hamilton, C. Horace, and Garnett, William E., "The Rôle of the Church in Rural Community Life in Virginia," Virginia Agricultural Experiment Station, *Bulletin 267*, 1929, p. 11.

especially in presenting the report of a community survey which is to be read by a great many people of differing interests and varying amounts of time available for studying the bulky text of the report. The charts not infrequently "sell" the survey and its recommendations.

For the purpose of showing certain general social facts about a state and indicating variations in different parts, a cartogram on a county basis is useful. Figure XLVI was constructed to show the percentage of church membership among the white population in each county of Virginia.

The variations are marked, ranging from less than 25 per cent of the population in 9 counties to 66 per cent or more in 7 counties. A technical criticism might be made of this chart on the ground that it shows too much. The percentages of "independent cities" are a little confusing, and also the figures boxed in the upper right-hand corner are not immediately clear. If only the map, the title, and the legend had been given, the import would have been obvious. A little study soon clarifies the meaning of the percentages, however. It is a chart which would make anyone interested in the church as a useful social institution stop and ask questions and ponder the meaning of the wide variations in apparent interest in the church.

Diagrammatic presentation of the plan of organization of an agency or institution is widely used. Governmental organizations, corporations, and social agencies are often complicated in their structure. It is almost impossible for anyone, even an official, to visualize the detail of the organization of a federal department, unless his imagination is aided by some graphic means. Also, it is difficult to grasp the ramifications and divisional relationships of a large city school system. But a chart sets these out clearly. Figure XLVII shows the organization of the attendance work of the New York City school system.[7]

9. STANDARD RULES OF GRAPHIC PRESENTATION

The Joint Committee on Standards of Graphic Presentation, composed of representatives of fifteen scientific societies and two government bureaus, worked out the more generally accepted rules for graphic presentation and published their report in the *Quarterly Publication of the American Statistical Association*, De-

[7] United States Children's Bureau, *Publication No. 17*,

1. The general arrangement of a diagram should proceed from left to right.

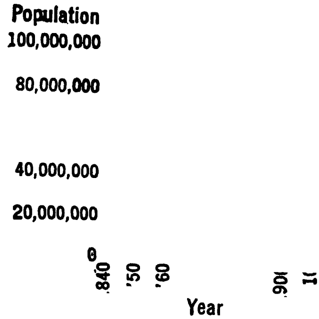


Illustration 1

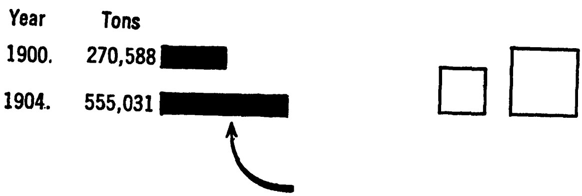


Illustration 2

2. Where possible represent quantities by linear magnitudes, as areas or volumes are more likely to be misinterpreted.

3. For a curve the vertical scale, whenever practicable, should be so selected that the zero line will appear on the diagram.

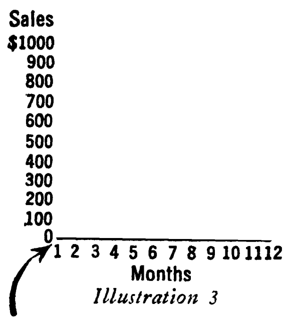


Illustration 3

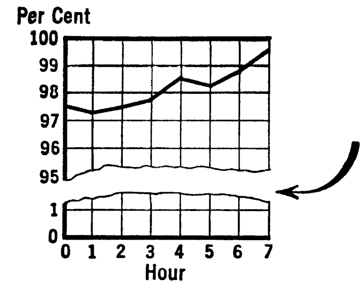


Illustration 4

4. If the zero line of the vertical scale will not normally appear on the curve diagram, the zero line should be shown by the use of a horizontal break in the diagram.

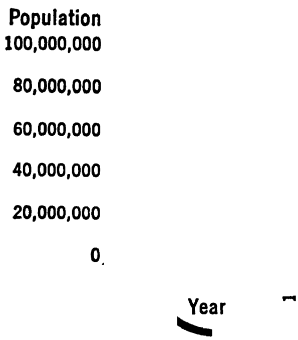


Illustration 5a

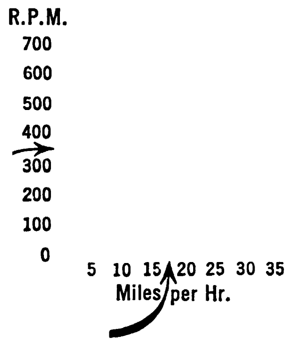


Illustration 5b

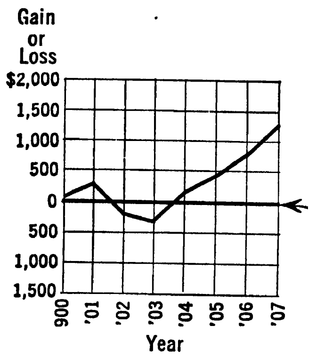


Illustration 5c

5. The zero lines of the scales for a curve should be sharply distinguished from the other coordinate lines.

6. For curves having a scale representing percentages, it is usually desirable to emphasize in some distinctive way the 100 per cent line or other line used as a basis of comparison.

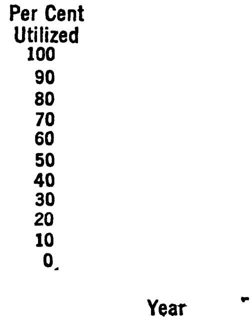


Illustration 6a

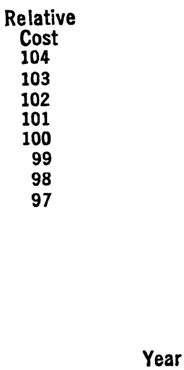


Illustration 6b

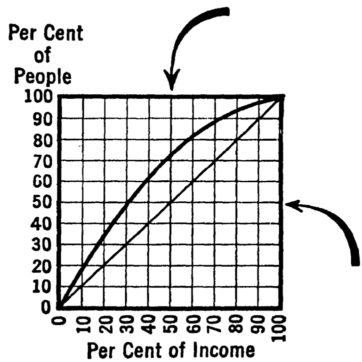


Illustration 6c

7. When the scale of a diagram refers to dates, and the period represented is not a complete unit, it is better not to emphasize the first and last ordinates, since such a diagram does not represent the beginning or end of time.

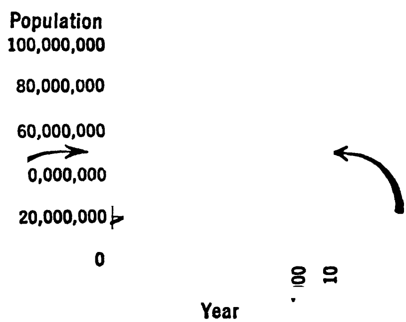


Illustration 7

Population
100,000,000

10,000,000

Year

Illustration 8

8. When curves are drawn on logarithmic coördinates, the limiting lines of the diagram should each be at some power of ten on the logarithmic scales.

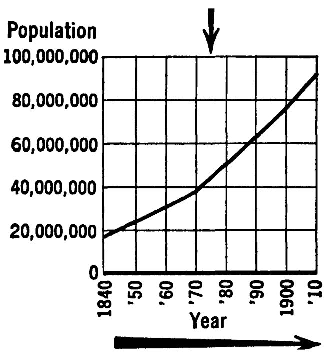


Illustration 9a

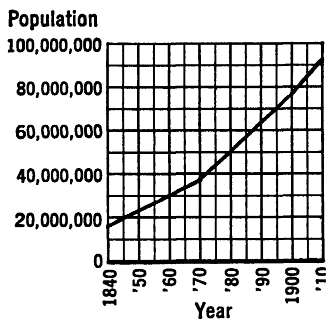
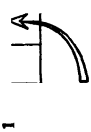


Illustration 9b

9. It is advisable not to show any more coördinate lines than necessary to guide the eye in reading the diagram.

Population
100,000,000
80,000,000
60,000,000
40,000,000
20,000,000
0



Year

Illustration 10

10. The curve lines of a diagram should be sharply distinguished from the ruling.

11. In curves representing a series of observations, it is advisable, whenever possible, to indicate clearly on the diagram all the points representing the separate observations.

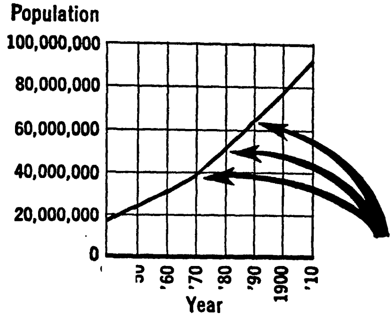


Illustration 11a

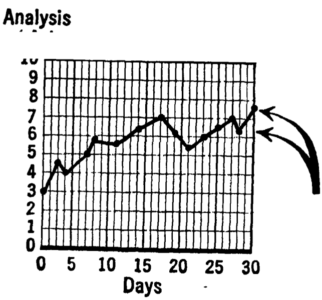


Illustration 11b

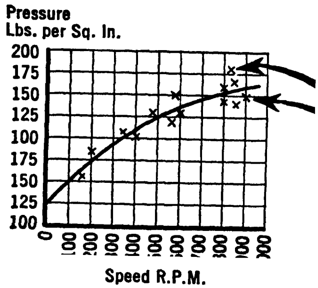


Illustration 11c

12. The horizontal scale for curves should usually read from left to right and the vertical scale from bottom to top.

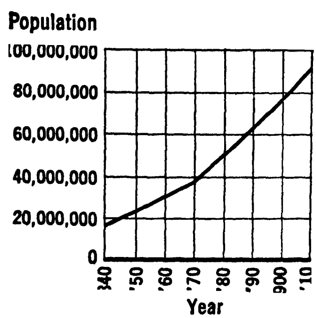


Illustration 12

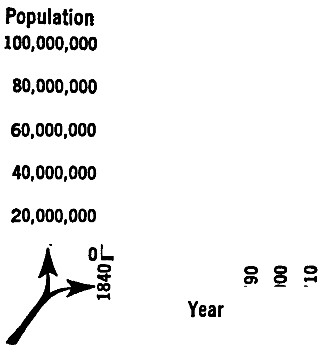


Illustration 13a

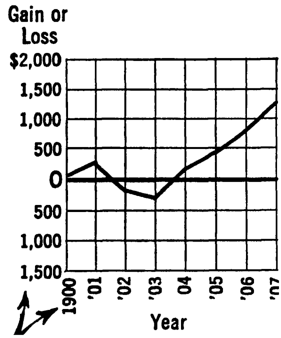
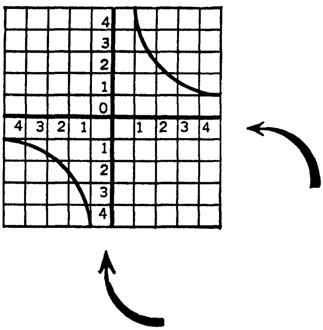
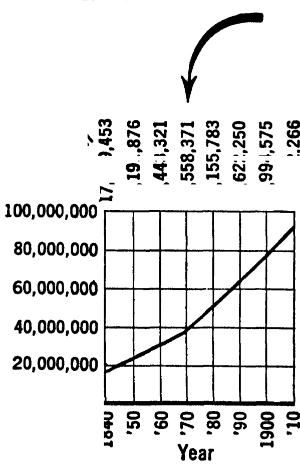


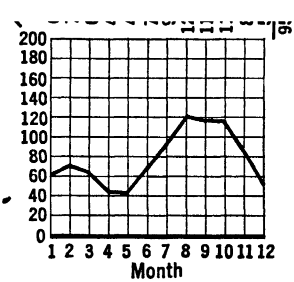
Illustration 13b



13. Figures for the scales of a diagram should be placed at the left and at the bottom or along the respective axes.

Illustration 13c





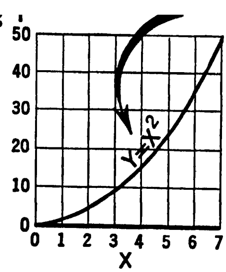
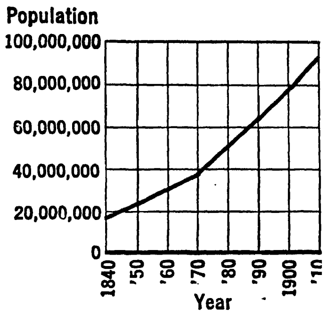


Illustration 14a Illustration 14b Illustration 14c

14. It is often desirable to include in the diagram the numerical data or formulæ represented.

15. If numerical data are not included in the diagram it is desirable to give the data in tabular form accompanying the diagram.



Year	Population
1840	17,069,453
1850	23,191,876
1860	31,443,321
1870	38,558,371
1880	50,155,783
1890	62,622,250
1900	75,994,575
1910	91,972,266

Illustration 15

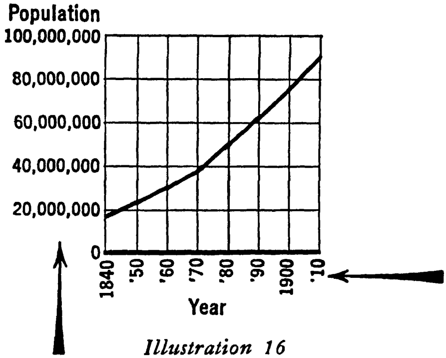


Illustration 16

16. All lettering and all figures on a diagram should be placed so as to be easily read from the base as the bottom, or from the left-hand edge of the diagram as the bottom.

17. The title of a diagram should be made as clear and complete as possible. Subtitles or descriptions should be added if necessary to insure clearness.

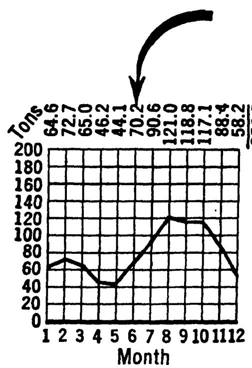


Illustration 17

Aluminum Castings Output of Plant No. 2, by Months, 1914.

Output is given in short tons.

Sales of Scrap Aluminum are not included.

cember, 1915. These rules have been quite generally used and are reproduced here for reference purposes.[8]

TABLE XXVI

WEIGHTED AGGREGATES OF PUBLIC WELFARE WORK AND THE ANNUAL TREND VALUES OF THE VOLUME OF WORK, INDIANA BOARD OF STATE CHARITIES, 1900 TO 1927[1]

Year	Weighted Aggregates	Annual Trend Values
Average	$126,917	$126,917
1900	$112,496	$110,583
1901	112,899	111,867
1902	112,560	113,151
1903	111,116	114,435
1904	111,656	115,719
1905	117,133	116,003
1906	117,133	117,287
1907	114,149	118,571
1908	121,505	119,855
1909	117,465	121,139
1910	118,345	122,423
1911	120,123	123,707
1912	124,344	124,991
1913	124,786	126,275
1914	132,039	127,559
1915	145,673	129,843
1916	140,469	131,127
1917	141,639	132,411
1918	126,688	133,695
1919	121,428	135,979
1920	117,109	137,263
1921	129,862	138,547
1922	136,027	139,831
1923	126,126	141,115
1924	135,807	142,399
1925	146,468	143,683
1926	152,806	144,967
1927	160,566	146,251

[1] Weighted aggregates are the sum of persons aided per 100,000 population by each Indiana agency or institution multiplied by the median cost per person for the respective agencies. Data from unpublished manuscript by the author.

[8] *Quarterly Publication of the American Statistical Association,* Vol. 14, pp. 790-797. The following scientific societies and government bureaus had representation on the Joint Committee: American Society of Mechanical Engineers, at whose invitation the Committee was formed, American Statistical Association, American Institute of Electrical Engineers, American Association for the Advancement of Science, American Academy of Political and Social Science, American Genetic Association, American Economic Association, United States Bureau of the Census, United States Bureau of Standards, American Association of Public Accountants, American Chemical Society, American Institute of Mining Engineering, American Psychological Association, Actuarial Society of America, and the Society for the Promotion of Engineering Education.

10. EXERCISES

1. Construct a straight line graph on the natural scale from the annual trend values given in Table XXVI.
2. Compute the annual growth of $1,000 at 6 per cent interest for a period of 10 years and construct a straight line graph on the natural scale for the annual amounts of the principal plus the simple interest accrued.
3. Tables XXVII and XXVIII give the population per square mile in continental United States, excluding Alaska, from 1790 to 1930 and patients per 100,000 population in Indiana hospitals for the insane on the last day of the fiscal year from 1900 to 1927. Plot these data:
   (a) on the natural scale;
   (b) on the semi-logarithmic scale. Explain the differences and the significance of each curve.
4. Tables XXIX and XXX give cumulative data. Make a chart from the data in each table and interpret the meaning of the charts.
5. Make bar charts representing the data in Tables XXXI and XXXII.

TABLE XXVII

POPULATION PER SQUARE MILE IN CONTINENTAL UNITED STATES, EXCLUDING ALASKA, 1790 TO 1930

Year	Population per Square Mile
1790	4.5
1800	6.1
1810	4.3
1820	5.5
1830	7.3
1840	9.7
1850	7.9
1860	10.6
1870	13.0
1880	16.9
1890	21.2
1900	25.6
1910	30.9
1920	35.5
1930	41.3

TABLE XXVIII

PATIENTS PER 100,000 POPULATION IN THE INDIANA HOSPITALS FOR THE INSANE ON THE LAST DAY OF THE FISCAL YEAR, 1900 TO 1927 [1]

Year	Patients per 100,000 Population
1900	173.4
1901	173.9
1902	179.9
1903	178.2
1904	188.3
1905	192.1
1906	195.9
1907	188.0
1908	188.9
1909	192.6
1910	198.0
1911	200.6
1912	212.3
1913	215.6
1914	216.9
1915	219.3
1916	219.9
1917	220.9
1918	207.4
1919	207.9
1920	206.6
1921	210.5
1922	217.3
1923	218.2
1924	217.3
1925	223.3
1926	226.2
1927	225.1

[1] From unpublished manuscript by the author.

TABLE XXIX

CUMULATIVE PERCENTAGES OF THE BUDGET ($72,000) EXPENDED BY A CHARITABLE AGENCY, FISCAL YEAR 1929–30, COMPARED WITH THE ESTIMATED AVERAGE MONTHLY BUDGET REQUIREMENTS

Month	Per Cent of Budget, Average Requirements	Per Cent of Budget, Expenditures in 1929–30
November	7.3	10.4
December	17.4	26.8
January	29.5	47.7
February	41.5	68.4
March	52.8	87.4
April	61.5	103.0
May	68.9	115.1
June	75.3	124.2
July	81.7	133.6
August	88.0	141.6
September	93.8	148.3
October	100.0	156.5

TABLE XXX

CUMULATIVE PERCENTAGES OF MALES IN THE POPULATION OF INDIANAPOLIS AND OF MALES EMPLOYED BY SIX INDIANAPOLIS FIRMS BY AGE GROUPS

Age	Per Cent of Males, City, 1920[1]	Per Cent of Males Employed by Six Firms, 1930[2]
15–19	10.3	4.8
20–24	23.0	23.7
25–29	36.5	42.2
30–34	48.9	57.7
35–39	61.2	70.8
40–44	71.0	80.5
45–49	80.2	87.2
50–54	87.8	92.8
55–59	92.9	96.6
60–64	96.8	98.8
65–69	100.0	100.0

[1] United States Census, 1920.
[2] Unpublished manuscript by the author.

TABLE XXXI

PERCENTAGE OF URBAN AND RURAL POPULATION IN THE UNITED STATES, 1890 TO 1930[1]

Year	Urban Population, Per Cent of Total	Rural Population, Per Cent of Total
1890	35.4	64.6
1900	40.0	60.0
1910	45.8	54.2
1920	51.4	48.6
1930	56.2	43.8

[1] United States Census, 1920, and *Population Bulletin, First Series*, 1931.

TABLE XXXII

PERCENTAGE OF TOTAL PERSONS RECEIVING POOR RELIEF IN POOR ASYLUMS AND FROM TOWNSHIP TRUSTEES (OUTDOOR RELIEF) IN INDIANA IN SPECIFIED YEARS[1]

Year	Poor Asylums, Per Cent of Total	Township Trustees, Per Cent of Total
1900	6.3	93.7
1905	6.4	93.6
1910	6.7	93.3
1915	3.4	96.6
1920	6.5	93.5
1925	4.6	95.4

[1] *Indiana Bulletin of Charities and Corrections, No. 182.*

6. Make sector charts representing the data in Tables XXXIII and XXXIV.

TABLE XXXIII

INMATES IN STATE PENAL AND CORRECTIONAL INSTITUTIONS PER 100,000 POPULATION, SEPTEMBER 30, 1929 [1]

	Number per 100,000 Population
Felons	126.6
Misdemeanants	41.7
Juveniles	25.9

[1] *Indiana Bulletin*, No. 182.

TABLE XXXIV

EXPENDITURES OF THE STATE GOVERNMENT OF NEW YORK BY GROUPS, PERCENTAGE GOING TO EACH, 1920 [1]

Group of State Expenditures	Percentage of Expenditures
All	100.1
Social	47.6
Protection	16.5
Administration	11.4
Construction	24.6

[1] Clark, Harold F., *The Cost of Government and the Support of Education*, p. 29. Teachers College, Columbia University, 1924.

7. Using data from the census of 1930, construct a cartogram of your state showing the percentage of foreign-born population in each county.
8. Using data from the census of 1930, construct
   (a) a histogram, and
   (b) a frequency curve of the age distribution of the population of the United States.

II. REFERENCES

Chaddock, Robert E., *Principles and Methods of Statistics*, Chap. XVI.

Lovitt, William V., and Holtzclaw, Henry F., *Statistics*, Chaps. V and VI.

Mills, Frederick C., *Statistical Methods*, Chap. II.

Mudgett, Bruce D., *Statistical Tables and Graphs*, Chaps. II and III.

Whipple, George C., *Vital Statistics*, Chap. II.

CHAPTER VIII

Measures of Central Tendency

1. INTRODUCTION

THE measure of central tendency, or the average, of a number of observations is probably the most commonly used method of statistical analysis. Almost any person with a common school education can think in terms of an average individual selected from a collection of similar individuals. But such a person may not know how to compute any measure of central tendency for the collection. Furthermore, it is necessary to know what kind of measure of the central tendency of the data is best for the purpose in mind. In colloquial language "average" is almost synonymous with "usual" or "most common." Among those who are familiar with statistical language, it generally means the arithmetic average. But there are several kinds of averages, and in order to emphasize the fact that they represent much the same thing, though differing in size and quality, the averages are referred to in this chapter as measures of central tendency—the tendency of the values of the individual items in any collection of data to cluster around some middle value.

As used in statistics, an average is a quantitative concept. It implies that some trait of the individuals can be measured and that an average value can be found for the separate values of this trait observed in individuals possessing it. In practice an average is computed for both variables and attributes, though strictly speaking the term should be used only in connection with true variables, that is, traits capable of being measured or counted. The central tendency of the values of the different items may be expressed by any of the averages, but in all cases it will be a quantity.

Another characteristic of an average is that it is a value typical of the data from which it is computed. It may be the most com-

mon value actually found, as the mode; or the middle value in a series arranged from lowest to highest, as the median; or it may be a value from which the minus deviations and plus deviations are equal, as the mean; or it may be a variation of the mean arrived at by taking the product of all the items and extracting the appropriate root, as the geometric mean. In any case it is a type value for the whole series. It can be used to represent the series in comparison with other type values of similar data. An average tells little about the individual items in a series; the actual variations of their values are disregarded, unless some method of relating variations to the average is used. Nevertheless, the typical value is useful as a shorthand description of the data, and is often an early step in much more complex statistical analysis.

The concept of central tendency is empirical. Experience with a great variety of facts has led to the inference that facts of the same kind differ in magnitude below and above a certain value with a fair degree of symmetry, and that a value may be found which is typical of the entire series. The low magnitudes differ from this value by about the same amounts as do the magnitudes above it. The leaves on a tree differ in length, but the extremely short ones and the extremely long ones are relatively rare. The heights of soldiers in a regiment differ, but the very short soldiers and the very tall ones are few in number as compared with the great majority. Some data are found to have the average among the low values or among the high values, because the distribution of values over the whole range is "skewed" one way or the other. For example, the average age of felons is fairly low, as compared with the average age of the total population. Consequently, there are some extremely high variations of age from the average age of felons, but that fact does not discredit the concept of central tendency; it merely suggests that some measure of variation from the average should be used in connection with it. There is a central tendency in the age distribution of felons. The central tendency of the magnitude of similar material facts is a matter of observation, and mathematics has provided a method by which this central tendency may be measured with a fair degree of accuracy. That is, the similarities of members of an animal species, the recurrent positions of a celestial body with reference to another celestial body, the usual age of marriage in a population, the usual number of children in dependent families, and the commonest level of intelligence found among delinquent boys were no-

ticed; and later statistical methods were used to determine the average magnitude of a trait of a species, the average observed position of the celestial body, the average age of males or females at marriage, the average number of children in dependent families, and the average intelligence of a group of delinquent boys. This is just another way of saying that the statistical method of averages is a means by which order is introduced into everyday experience and ascertained results are substituted for impressions.

An average is most significant when the data have a high degree of homogeneity. This is particularly to be emphasized in dealing with social data, because so many factors affect a datum to render it highly variant from other data of perhaps the same general type. Age, sex, nationality, and race are factors which must be considered in connection with the study of some other social factor, such as crime, because they may lower the degree of homogeneity and render an average of any sort meaningless. For example, in the study of crime the results are more dependable if juvenile delinquents are studied separately and if the sexes are studied separately. Where one or more nationalities enter into the situation, it is usually desirable to consider them separately for the purpose of determining the nationality having the highest or lowest rate of crime. If a study of wages in a factory is being made, the study should be divided into analyses of wages for each sex, and wages for office workers and industrial workers. It so happens that by custom female workers get relatively less pay for similar work than male workers, and the office wage scale is generally quite different from the plant wage scale. Homogeneity of data is increased when such divisions of workers are made. The intelligence quotient of an individual is affected by the social class from which he comes. If he is taken out of a low social class and put into a higher social class, his intelligence quotient frequently rises, or, if a better social adjustment within his own social class is made, his intelligence quotient is known to rise. The average intelligence of all the children in a given school might have some meaning, but, if the children could be separated into the social classes to which they belong and the average intelligence of each social class obtained, the several averages would vary considerably, reflecting the heterogeneity of the total school population and showing the greater significance of average intelligence when it refers to one fairly definite social class. When primary data are to be collected

or secondary data are to be used, early consideration must be given to their homogeneity.

Five averages are usually recognized. They are the mode, the median, the arithmetic mean, the geometric mean, and the harmonic mean. This chapter discusses all of these except the harmonic mean, but omits the latter because it is not often used. It is customary in books on statistics to discuss the averages in the following order: arithmetic mean, median, mode, and geometric mean, though some variations in the order do occur. Although the arithmetic mean is the form of average most used, it is not the concept most people have in mind when they use the term "average." What they think of is the "usual" magnitude of a factor, and this is the concept of the mode. Hence, pedagogically it seems more appropriate to discuss the mode first, and then follow it with a discussion of the median, which resembles the former in one respect, namely, that it is also a position average. The arithmetic mean and the geometric mean obviously belong together, rather than the arithmetic mean and the median, and of the two means the geometric is less well known and less useful. For these reasons, the order of discussing the averages will be as follows: mode, median, arithmetic mean, and geometric mean.

Before turning to the methods of computing the averages, attention should be called to the variation of values around an average, otherwise known as deviation from the average, or dispersion. Too great emphasis should not be placed upon the significance of an average, unless some measure of dispersion is used along with it. If the dispersion is small, the inference to be drawn is that the homogeneity of the data is high and the average reliable; on the other hand, if the dispersion is great, homogeneity is low and the reliability of the average doubtful. Students of the social sciences and of social work are interested not only in the central tendency of a body of data but also in the dispersion of the individual values around the central tendency. Measures of dispersion will be discussed in Chapter IX, but it is important for the student to realize at this point that an average requires these checks to make its significance clear.

2. THE MODE

The mode is the most common value occurring in a collection of data. It is the rule-of-thumb average which corresponds to the concept in the mind of the person untrained in statistics. The con-

cept may be made clearer, if the mode is referred to as the "fashion." A fashion refers to the dominant trend of a certain kind of

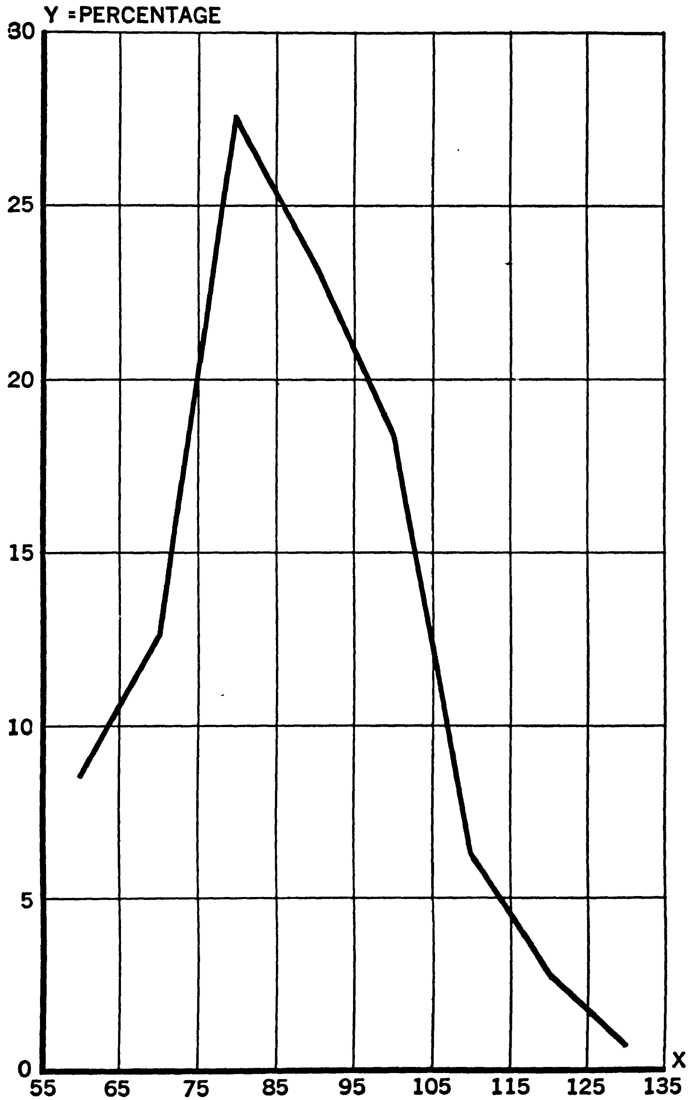


FIGURE XLVIII.—DISTRIBUTION OF INTELLIGENCE AMONG 451 CHILDREN IN DEPENDENT FAMILIES

behavior, such as wearing empire hats or explaining behavior in terms of psychoanalysis. The difference between the statistical con-

cept of the mode and the concept of the fashion of the day lies in the fact that the mode is quantitative while fashion to a large extent is a qualitative concept, though in some instances it might be reduced to quantitative definition. Figure XLVIII will illustrate the mode, using I.Q's of children in dependent families. The highest point of the curve, which is on the ordinate at the mid-point of the class-interval marked I.Q. 75-85, indicates the mode. The value of the mode is in this class-interval, though the exact value of it cannot be determined from an examination of the diagram. Twenty-seven and six-tenths per cent of all the children had I.Q's between 75 and 85, whereas by the same test the largest grouping of I.Q's would theoretically be between 95 and 105. This difference between the mode for the children in dependent families and those in an unselected group shows that children in dependent families rate lower in intelligence tests because of natural inferiority or because of social conditions. The modal class-interval enables us to make this comparison with the unselected group. The "average" dependent child, as measured by the mode, had an I.Q. between 75 and 85.

The above method of determining the mode is known as inspection. But inspection may be used in other ways of finding the mode. The simplest method is the array. The items are arranged in ascending order from the lowest value to the highest, as shown in the following table:

TABLE XXXV

An Array of the Ages of 100 Felons Selected at Random from Cases Disposed of by the Marion County, Indiana, Criminal Court in 1930

Age	Age	Age	Age
16	18	20	22
16	18	20	22
17	18	20	22
17	19	20	22
17	19	20	22
17	19	20	22
17	19	20	23
17	19	20	23
17	19	21	23
17	19	21	23
18	19	21	23
18	19	21	24
18	19	21	24
18	20	22	25
18	20	22	25
18	20	22	25

TABLE XXXV—(*Continued*)

Age	Age	Age	Age
26	30	35	45
26	31	36	46
27	32	38	47
28	32	39	48
28	33	41	49
28	33	43	49
29	34	43	54
30	34	43	55
30	34	45	55

The age group that is the most numerous is the modal group. In this array more felons are 20 than any other age. If these ages were presented as a frequency distribution in diagram form, it could easily be seen that the 20-year group is the largest; that is, this age is the most usual age of felons in this group. A larger number of cases, however, might show that mode to be located in some other age group.

The mode may be determined roughly from grouped data by successive regrouping. Table XXXVI illustrates this method:

TABLE XXXVI

LOCATION OF THE MODE BY SUCCESSIVE REGROUPING OF AGES OF FELONS

Age 2-Year Interval		Four-Year Group		Six-Year Group	
	f	4-Year Interval f	Shift One Interval f	6-Year Interval	Shift One Interval f
(1)	(2)	(3)	(4)	(5)	(6)
16–17	10	29	(Omit 10)		(Omit 10)
18–19	19			45	
20–21	16	30	35		49
22–23	14				
24–25	5	8	19	22	
26–27	3		7		12
28–29	4	8			
30–31	4		8	12	
32–33	4	8			12
34–35	4				

TABLE XXXVI—(*Continued*)

Age 2-Year Interval		Four-Year Group		Six-Year Group	
		4-Year Interval	Shift One Interval	6-Year Interval	Shift One Interval
(1)	f (2)	f (3)	f (4)	f (5)	f (6)
36–37	1	3		7	
38–39	2		3		4
40–41	1	4			
42–43	3		5	6	
44–45	2	4			7
46–47	2		5		
48–49	3	3		5	
50–51	0		0		3
52–53	0	3			
54–55	3		(Omit 3)	3	(Omit 3)

In each column of frequencies the largest is underlined to indicate the modal age group. It will be noticed that the size of the class-interval causes the modal group to shift. With a two-year interval the mode falls in the 18-19 year class, but in the four-year interval it falls in the 20-23 year class. In the six-year interval it falls in the 16-21 year class. When the first and last class frequencies are omitted, as in columns (4) and (6), the mode is between 18 and 21 years and 18 and 23 years, respectively. Because the mode is a position average, the omission of a few extreme items should not, and in fact does not, affect it. The class-interval 18-19 appears in four out of five of the groupings, and the class-interval 20-21 appears in four. This would seem to indicate that the mode lies between these two groups, or at about 20 years, the latter being the mode determined from the array. The method of successive regrouping gives what Chaddock has called the crude mode.

The mode may be computed from grouped data by means of the following formula:

$$Mo = l +$$

Mo = the mode
l = lower limit of the class-interval having the largest number of frequencies

f_1 = number of items in the class just below the modal group
f_2 = number of items in the class just above the modal group
i = size of the class-interval

Using the two-year class-interval as a basis of computing the mode for data in Table XXXVI, the following substitutions in the formula are made:

$$Mo = 18 + \frac{16}{16 + 10}\,2$$
$$= 19.23 \text{ years}$$

If a four-year class-interval is used, the mode is 20.07. When the length of the class-interval changes, the number of items in each class varies also, and the effect is to shift the mode slightly.

Pearson has suggested another formula for ascertaining the mode for data distributed in the form of a bell-shaped curve or only moderately skewed to the right or left. The formula is as follows:

$$Mo = \text{Mean} - 3\ (\text{Mean} - \text{Median})$$

In a perfectly bell-shaped curve, the three measures of central tendency are identical, but in moderately asymmetrical distributions they differ by amounts among which there is a fairly constant relationship. This formula could not be applied to the felony data used above, because the age distribution is skewed far to the right in the direction of the higher ages. The constant relation required for the application of this formula among the mean, median, and mode does not exist in such highly asymmetrical distributions as ages of felons.

All the methods for locating the mode so far discussed give approximations to it, but they do not give it exactly. The only exact method is to fit an ideal frequency curve to the actual figures.[1] This method is complicated and is beyond the scope of the discussion at this point. Furthermore, the limited use to which the mode may be put does not often warrant the laborious calculations required to obtain it exactly. Methods giving approximations, as discussed above, are all that the student will ordinarily require in social statistics. The methods of arriving at the crude mode may be used with either continuous or discontinuous (discrete) data, but the exact method should be used only with continuous data.

The mode has some advantages as a rough measure of central tendency. It marks the approximate location of the most common value in a series of data. And this may have practical significance.

[1] Yule, *op. cit.*, p. 121.

It may be important to a court to know that the modal age of felons it has sentenced is about 20 years, whereas the mean age is considerably higher. In the study of wages it is sometimes important to know in what wage class the mode falls. Another advantage is that, being a position average, the mode is not affected by the addition or subtraction of an extreme item any more than it is by the addition or subtraction of an item near its own value. A third advantage is that the skewness of a distribution may be measured in terms of the mode. But the mode has its limitations as an average: It is affected more than the mean by changes in the length of the class-interval; it cannot be exactly computed without resort to complicated and laborious methods of curve-fitting; it does not lend itself to the algebraic treatment that may be required in further statistical analysis; it cannot be used in connection with time series, because the high points on such a curve represent abnormal and not modal conditions. The student should study his data carefully. If it appears that the mode is really a significant measure of the data, he should determine the mode and make use of it in his interpretation. Whether or not the mode is useful in a given case will depend upon the data themselves and upon the purpose of the investigator. The use of the mode as a method of statistical analysis is not a purely mechanical matter; it is a means to an end, and, unless determination of the mode throws light on the problem, there is no point to finding it.

3. THE MEDIAN

The median is the middle value in a series of data, when they are arranged in ascending order from lowest to highest. It may fall on an actual item in the series or it may lie between two items. Like the mode, it is a position average and is not seriously affected by the addition or subtraction of an item, large or small. In a series of data the median is the value above and below which the numbers of items are equal. The chances are even that, if an item is selected at random from the series, it will be greater or less than the median.

The median may be found for both ungrouped and grouped data. If their number is small, the items can be arrayed and the median found by inspection. In any series the first step is to locate the position of the middle value. Take the following numbers: 4, 6, 7, 9, 11, 14, 15. They are odd, and the median value is 9. But if another item is added above 15, the median value then falls between 9 and 11. Referring to Table XXXV, how can the median

position be located? This table contains an even number of items: 100. The median value lies between the 50th and 51st items. For an even series, the median position may be located by this simple formula:

$$2$$

$$= \frac{100 + 1}{2}$$

$$= 50.5\text{, the median position}$$

$$Md_p = \text{the median position}$$

$$N = \text{number of items in the series}$$

In case of an odd number of items, the same formula is used, but the median position will be the position of the item standing halfway between the two extreme items, and will be a real item. In the arrayed items of Table XXXV, the median lies between the 50th and 51st items, both of which happen to be 22 years. Hence, the median by inspection is found to be 22 years. If the median position had been between two values of unequal size, a further step would be necessary. Suppose the median had fallen between 22 and 23 years. Then the procedure would be to add the two values and divide by 2 which would give 22.5.

When a series contains a large number of items, the data are generally grouped into class-intervals. The following table shows the age distribution of male workers in Boston who were unemployed at the time of the census of unemployment in 1930.

TABLE XXXVII

UNEMPLOYED MALE WORKERS IN BOSTON BY AGE GROUPS, APRIL, 1930[1]

Age	Number of Workers
Total	21,262
10–14 years	13
15–19 "	1,745
20–24 "	2,968
25–29 "	2,448
30–34 "	2,176
35–39 "	2,323
40–44 "	2,234
45–49 "	2,195
50–54 "	1,786
55–59 "	1,544
60–64 "	1,107
65–69 "	723

[1] *Unemployment Bulletin*, Massachusetts, United States Bureau of the Census, p. 12.

The general formula for the computation of the median from grouped data is:

$$Md = l + \frac{\frac{\quad}{2} - F}{f} i$$

Md = the median
l = value of the lower limit of the class-interval which contains the median
I = total number of items plus one
F = sum of all frequencies in classes below l
f = number of items in the class-interval containing the median
i = value of the class-interval

Substituting in the formula to find the median for the data in Table XXXVII, we have:

$$= 35 + \frac{\frac{21262 + 1}{2} - 9350}{2323} 5$$

$$= 35 + 2.8$$
$$= 37.8, \text{ the median}$$

The median is found to be 37.8 years—that is, correct to one decimal place. As reported by the census there were a few unemployed persons in the "unknown" class and a few in the class of "70 years and over." These were omitted from the table from which the above median was computed. Those whose ages were known to be 70 or over could have been included in the table, but the unknown group could not be used, because there is no way of distributing this group among the established class-intervals.

The student will have noticed that the formula utilizes the class-intervals and frequencies below the median. As a matter of fact, the median may be computed by using the class-intervals and the frequencies above it in a similar manner. Changing certain letters in the formula for purposes of clarity, and changing the first plus sign to a minus, the formula for using the upper half of the data would be as follows:

$$Md = L - \frac{\frac{N + 1}{2} - F_1}{f} i$$

F_1 = sum of all items in classes above L
L = value of the upper limit of the class-interval containing the median

Substituting in this formula:

$$Md = 39{\cdot}99 - \frac{\frac{21262 + 1}{2} - 9589}{2323} 5$$

$$= 39{\cdot}99 - \frac{1042.5}{2323} 5$$

$$= 37{\cdot}75, \text{ the median}$$

Note that in the table the class-interval containing the median is indicated to be 34.39. That means that all ages of 34 and *less* than 40 are included in this class-interval. When computing the median from the top down, it is necessary to express the upper limit more exactly than the figure 39. Hence, it is indicated here to be 39.99, and the median by this method is 37.75, or five-hundredths less than the median by the previous computation, but this difference is due only to the fact that the upper limit of the class-interval was expressed to the hundredth of a year. It might have been expressed to the ten-thousandth of a year, in which case the difference between the first and second methods would have been five ten-thousandths. The important point is that, for all practical purposes, the medians are the same. It is more common, however, to find the median computed from the bottom upwards.

When the data are grouped the median may also be located with approximate accuracy by graphic methods. The most common method for doing this is the use of "less than" and "more than" cumulative frequency curves drawn on the same paper. Table XXXVIII gives the cumulative frequencies and Figure XLIX locates the median graphically below the point of intersection. The exact value of the median cannot be determined from this graph, but apparently it is about the same as the computed median, that is, 37.8. The two cumulative frequency curves intersect at a point which divides the total frequencies in half and, consequently, at a point whose value is the middle value of the series. It should be noticed that the curves must be plotted, not at the mid-points of the spaces representing age periods, but on the ordinate representing the lower limit of the class-interval. If they are plotted at the mid-points of the spaces, the intersection of the two curves

TABLE XXXVIII

CUMULATIVE FREQUENCIES, UNEMPLOYED MALE WORKERS IN BOSTON

Age in Years	*f*	(I) "Less Than" Number	(II) "At or More Than" Number
10–14	13	0	21,262
14–19	1,745	13	21,249
20–24	2,968	1,758	19,504
25–29	2,448	4,726	16,536
30–34	2,176	7,174	14,088
35–39	2,323	9,350	11,912
40–44	2,234	11,673	9,589
45–49	2,195	13,907	7,355
50–54	1,786	16,102	5,160
55–59	1,544	17,888	3,374
60–64	1,107	19,432	1,830
65–69	723	20,539	723
		21,262	0

is to the right of the middle value; that is, the value indicated is too high. The chief value of the graphic method of locating the median is for interpretation to persons reading a report or listening to one which may be made orally. A chart can be presented which gives proper perspective and impresses the observer immediately as to the value of the median. If the numerical value is all that is wanted, it is easier to compute the median by formula.

Quartiles, deciles, and percentiles are frequently discussed in connection with the median, because they are associated with it. But these measures are not, in fact, measures of central tendency, but of dispersion. Kelley recognizes this, though he includes them in his chapter on measures of central tendency.[2] Secrist uses the concept of these methods which is utilized here and discusses it in his chapter on dispersion.[3] There is no more reason for the discussion of quartiles, deciles, and percentiles in juxtaposition to the discussion of the median than there is for the discussion of the standard deviation on the same page with discussion of the mean. These methods will be considered in the next chapter.

As an average the median has at least two advantages not equally possessed by other averages: (1) it is easily calculated, and (2) it is not significantly affected by a few extreme items in a series. The median has been widely used in anthropometric and

[2] Kelley, Truman L., *Statistical Method,* p. 59. New York: Macmillan, 1923.

[3] Secrist, Horace, *An Introduction to Statistical Method,* Chap. X. New York: Macmillan, 1929.

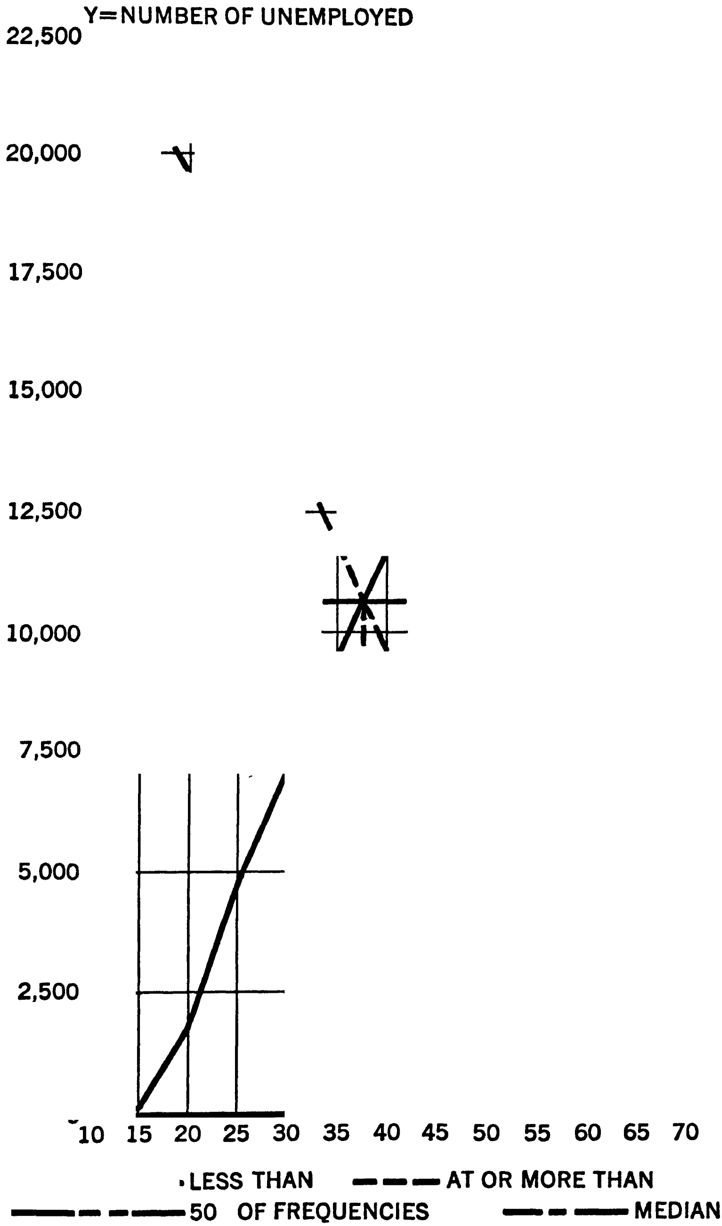


FIGURE XLIX.—LOCATION OF THE MEDIAN BY MEANS OF CUMULATIVE FREQUENCY CURVES

educational measurements to locate the point above and below which 50 per cent of the items lie. In establishing norms of distribution of traits the median value of the series is frequently important. But the median should not be used without careful consideration of other questions. If the frequency distribution with which the student is working is bi-modal, the median may fall at a point not representative of the series. It may be unrepresentative in a series with a single mode, but caution in its use is particularly important if the series shows two or more modal points. The median does not lend itself to numerical and algebraic treatment; the algebraic sum of the deviations of the individual items from the median is not zero. Sometimes the computation of an average is only one step in a problem requiring statistical analysis, in which case it may be necessary to choose an average, such as the mean, which lends itself to algebraic uses.

4. THE ARITHMETIC MEAN

The arithmetic mean is a measure of central tendency derived from consideration of all the values in the series. It is affected by the size of every item included in the computation. When the word "mean" or "average" is used without qualification, the arithmetic mean is usually meant. Among statisticians it is the most frequently used average. Yule[4] points out that the arithmetic mean fulfills more of the conditions of an average than does any other measure of central tendency. He names six conditions: (1) an average should be rigidly defined; (2) it should be based upon all observations; (3) it should be readily comprehensible; (4) it should be easily and rapidly calculated; (5) it should be as little affected by fluctuations of sampling as possible; and (6) it should lend itself readily to algebraic treatment. The arithmetic mean fulfills all these conditions, except (5) which may not be fulfilled if there is a number of extremely small or extremely large items. The median would be less affected under such conditions.

The mean may be computed from either ungrouped or grouped data, but the methods are somewhat different. Referring to Table XXXV, the mean age of 100 felons would be the sum of the ages divided by 100. The formula is:

M = the mean
X = the individual item in the series

[4] *Op. cit.*, pp. 108, 109, 119, 120.

Σ = "the sum of" the individual items
N = number of items

Hence,

100

= 26.37, the mean

This is the absolute mean and is the one commonly thought of when the mean is mentioned. If only a small number of items is involved, the sum of the individual items may be easily obtained, but, if the items happen to run into the thousands, the work would be considerable. Therefore, it is desirable to have a method of computing the mean from data grouped into class-intervals.

To illustrate the method of computing the mean from grouped data we shall use the data for unemployed workers in Boston as given in Table XXXVII. The formula differs slightly from that for computing the mean for ungrouped data. It is as follows:

$$N$$

in which m is the mid-point of the class-interval and f the number of items within the class-interval. The other symbols have the same meaning as in the previous formula. In order to compute the mean from grouped data it is necessary to set up a table. The tabular form makes the process clearer and enables the student more easily to check the accuracy of his work. To compute the mean age of unemployed workers in Boston, the following table is given:

TABLE XXXIX

Computation of the Mean by the Long Method for Grouped Data: Unemployed Workers in Boston, Total 21,262

Age Years (1)	Mid-Point of Class-Interval m (2)	Number of Unemployed f (3)	Product of Columns (2) and (3) fm (4)
10–14	12.5	13	162.5
15–19	17.5	1,745	30,537.5
20–24	22.5	2,968	66,780.0
25–29	27.5	2,448	67,320.0
30–34	32.5	2,176	70,720.0
35–39	37.5	2,323	87,112.5
40–44	42.5	2,234	94,945.0
45–49	47.5	2,195	104,262.5
50–54	52.5	1,786	93,765.0
55–59	57.5	1,544	88,780.0
60–64	62.5	1,107	69,187.5
65–69	67.5	723	48,802.5
Total		21,262	822,375.0

Substituting in the formula, we have:

$$\frac{\quad}{21{,}262} = 38.68 \text{ years, the mean}$$

The use of this table and formula shorten the work for computing a mean for 21,262 items. It could be done by adding all the separate items and dividing by 21,262, but the work would be much greater. If the data are given in a frequency table, the mean could not be computed by adding the separate frequencies and dividing by the total number. It is, therefore, necessary to have a method of computing the mean from grouped data. But one caution should be kept in mind when computing a mean by this method. The mid-point of each class-interval was multiplied by the number of items in the class-interval. For example, in the first class-interval of the table the mid-point is 12.5, and this number is multiplied by 13. It is assumed that some of the ages are less than 12.5 years and that some are greater but that the sum of the differences of those less than 12.5 is equal to the sum of the differences of those above 12.5. We have assumed an even distribution of the items throughout the class-interval. This assumption is made regarding each class-interval in the table. If for any reason it were likely that the items in each class-interval tended to concentrate at either the lower or the upper end of the class-interval, the computed mean would probably be erroneous, because the assumption of even distribution would be unjustified. It was pointed out in Chapter VI[5] that uneven distributions do occur. The student should consider carefully whether or not his frequency distribution is of this sort. If it is suspected that the data in a frequency distribution may have this tendency and there is no way of rearranging the class-intervals because of the absence of the original data, then some reservation is in order as to whether the computed mean is exact or not.[6] That reservation may properly be made regarding the mean age of unemployed workers above, because ages, even in the United States census, have been known to show some concentration around ages divisible by 5. No redistribution can be made of the class-intervals which would test this fact,

[5] See p. 148.

[6] Sheppard has suggested a correction for the standard deviation of a distribution characterized by unevenness within the class-interval. In such cases the true standard deviation is: $\sigma^2 = \sigma_1^2 - \frac{c^2}{12}$. See Yule, *op. cit.*, pp. 211, 212.

because the original data are not published by the census; it would not be possible to put the ages divisible by 5 at the mid-point of class-intervals. Hence, it may be that 38.68 years is not the exact mean but only an approximation to it.

Although the preceding method of computing the mean saves time as compared with the method of adding the separate items and dividing by the number of items, it is still a "long method" for computing the mean. It involves the use of large numbers and long multiplications, and when large numbers are used the chance of mistakes is increased. The statistician should employ all possible methods to eliminate opportunity for errors. There is a shorter method, sometimes called the "deviation method," of computing the mean.

The algebraic sum of the deviations from the mean is zero. This fact may be used to compute the mean by a "short method." The student may take any small number of items for which he has computed a mean, express the deviations of each item from the mean with their appropriate signs, and add the deviations algebraically. The result will be zero.

Since we know that the sum of the deviations from the mean is equal to zero, we can take any arbitrary origin in the frequency distribution, assume the mid-point represented by this origin to be the mean, compute a correction factor, add the correction factor to the assumed mean, and the result is the true mean of the frequency distribution. This principle will be illustrated by the same data as were used to illustrate the computation of the mean by the long method.[7]

The assumed mean in Table XL is 37.5 years. That becomes the arbitrary origin from which to measure step-deviations, that is, deviations from the mid-point of the class-interval containing the assumed mean expressed in class-interval units. In column (4) the arbitrary origin is marked 0 and the step-deviations of class-intervals whose values are lower than that of the class in which the mean is assumed to be are marked minus. The step-deviations of class-intervals higher in value than the origin have a plus sign, but, following conventional procedure, the sign is not expressed.[8] The frequencies, f, are multiplied by their respective step-deviations, d, and the product of any f and any d takes the

[7] For more detailed proof of the short method for computing the mean, see Mills, Frederick C., *Statistical Methods*. New York: Henry Holt & Co., 1924.

[8] In this book, wherever the algebraic sign in front of a quantity is unexpressed it is plus.

TABLE XL

COMPUTATION OF THE MEAN BY THE SHORT METHOD

Age (years)	Mid-Point of Class-Interval	Frequency	Deviations from Assumed Mean in Class-Interval Units	fd	
	m	f	d	−	+
(1)	(2)	(3)	(4)	(5)	(6)
10–14	12.5	13	−5	65	
15–19	17.5	1,745	−4	6,980	
20–24	22.5	2,968	−3	8,904	
25–29	27.5	2,448	−2	4,896	
30–34	32.5	2,176	−1	2,176	
35–39	37.5	2,323	0		
40–44	42.5	2,234	1		2,234
45–49	47.5	2,195	2		4,390
50–54	52.5	1,786	3		5,358
55–59	57.5	1,544	4		6,176
60–64	62.5	1,107	5		5,535
65–69	67.5	723	6		4,338
		21,262		23,021	28,031

sign of the *d*. For convenience two columns are provided in the table, one for the —*fd's* and the other for the +*fd's*. The totals of columns (3), (5), and (6) are obtained. Now we are ready to substitute in the formula for the short method, which is:

$$M = \quad + c$$

in which M is the mean, M_1 is the assumed mean, or 37.5 in this case, and c is the correction factor.

$$= \frac{\Sigma fd}{N}, \text{ in steps or class-intervals}$$

$$c = \frac{\quad}{N} i, \text{ in years}$$

$$= 38.68 \text{ years}$$

In order to reduce the correction factor to terms of years, it must be multiplied by the size of the class-interval, 5. This is indicated above by the symbol, *i*. It should be noted that the correction factor is *added* in the algebraic sense; that is, with due regard to signs. If the assumed mean is higher than the true mean, the correction factor will be a minus quantity. The mean computed by this short method is exactly the same as by the long method.

In order to test whether or not the results are the same, when different arbitrary origins are used, we might try several others. The author has computed the mean from 47.5 as assumed mean, and the result is 38.68 years, the same as the preceding result. The result will be the same, regardless of what the assumed mean is, but, if the assumed mean is taken near the true mean, the figures dealt with are smaller and, consequently, more readily handled.

The same caution as to the concentration of values at some point in the class-interval holds for the short method as for the long method. The class-interval should not be too large, and, if it is known that concentration of the values of items occurs at a certain place, this point should when possible be placed in the middle of the class-interval.

There is another variation of the mean, but it is still an arithmetic mean. That is the *weighted mean.* The method of computing the weighted mean resembles the method of computing an ordinary mean from a frequency distribution, but in practice the concept is more restricted and should not be applied to a frequency distribution. The concept should be used in connection with a mean computed from rates or ratios, and it is widely used in the construction of index numbers. The formula is:

Computation of a weighted mean will be illustrated from a series of index numbers for various types of public welfare work in Indiana. These indexes are based upon the number of clients per 100,000 population of the state who were under the care of the agencies at the end of the fiscal year, 1930.

$$\frac{12{,}983.30}{100.0}$$

$$129.83$$

The weighted mean for these index numbers is 129.83. The percentages in column (2) represent changes for the same institutions and agencies from the numbers of people they were serving in 1913 (in each case the number served in 1913 is taken as 100 per cent). Consequently, the weighted mean shows that the same

TABLE XLI

COMPUTATION OF THE WEIGHTED MEAN INDEX NUMBER FOR THE NUMBER OF CLIENTS UNDER THE CARE OF PUBLIC WELFARE AGENCIES IN INDIANA, SEPTEMBER 30, 1930. BASE, 1913

Agency (1)	Index Number X (2)	Weight—Per Cent of Total Clients W (3)	Col. (2) × Col. (3) WX (4)
Total	1,925.4	100.0	12,983.30
Hospitals for Insane	97.9	23.8	2,330.02
School for Feeble-minded Youth	107.7	5.4	581.58
Colony for Epileptics	307.6	2.7	830.52
Soldiers' Home	32.5	1.2	39.00
Soldiers' and Sailors' Orphans Home	117.3	2.2	258.06
Tuberculosis Sanatorium	124.4	.6	74.64
School for the Deaf	116.8	1.4	163.52
School for the Blind	86.9	.5	43.45
State Prison	168.7	8.1	1,366.47
Reformatory	177.9	6.9	1,227.51
Women's Prison	131.4	.7	91.98
Boys' School	80.0	1.7	136.00
Girls' School	113.6	1.3	147.68
Poor Asylums	133.5	16.9	2,256.15
Dependent and Neglected Children, Wards	129.2	26.6	3,436.72

agencies were caring for 29.83 per cent more persons in 1930 in proportion to population of the state than in 1913. The simple mean of the index numbers is 128.40, or more than 1 per cent less than is shown by the weighted mean. This is not a large difference, but in some cases the weighted mean may vary much more from the simple mean. The weighted mean is used extensively in finding the average price of a commodity on a certain day. For example, the price of eggs of the same grade will vary in price among a number of stores, and variations among different grades will be still larger. The only way to arrive at a figure which fairly expresses the general price of eggs in a city on a given day is to weight the price of different grades and of prices for like grades at different stores by the quantities sold on that day. In constructing an index number of the cost of living, the United States Bureau of Labor Statistics made extended studies of the quantities of different articles used in the family budgets of a large sample of families. Weights were determined on the basis of the quantities used, and then the average prices of the commodities were multiplied by the weights to give proper importance to each

item. An index number which would fairly represent the cost of living could not be computed without using a system of weights, based upon the relative importance of the items included.

The quantity used for a weight is to a considerable extent arbitrary. Whatever it is, it represents the worker's estimate of the relative importance of the items which are to enter into the weighted mean. Pounds, inches, dollars, ratios, etc., may constitute the weights. Because of the arbitrary element in weighting, it is sometimes said that better results would be obtained by neglecting weights. But this is obviously fallacious, because differential importance is given to the items in a series of rates or ratios regardless of the presence or absence of a weighting plan. For example, the index number for patients in the Indiana Colony for Epileptics was 307.6 in 1930, whereas the index number for patients in the hospitals for the insane was 97.9. The number of patients in the Colony for Epileptics was 767 at the end of the fiscal year, 1930, while the number of patients in the hospitals for the insane at the same time was 6,839. The rate of increase for the Colony for Epileptics, compared with 1913, was very large, whereas the hospitals for the insane show a slight decline. Whether or not a system of weights is used, there is weighting—that is, the rate of change in the population of the Colony for Epileptics is given an importance which in fact it does not have. If instead of finding the mean index number for the 15 public welfare agencies and institutions, it was desired to find only the mean index number for the hospitals for the insane and the Colony for Epileptics, the importance of weighting is made still clearer. The simple mean of 97.9 and 307.6 is 202.8, which implies a doubling of the number of patients in the seven institutions represented since 1913. But the population of the hospitals for the insane has actually declined slightly relative to population in Indiana; it is only the population of the Colony for Epileptics that has shown a rapid increase, and the number of patients in the Colony for Epileptics in 1913 was small. If new percentage weights are computed for epileptics and insane only, and the index numbers are multiplied by these, the weighted mean index of population of these two types of public welfare institutions is 119.1, as compared with an unweighted mean of 202.8. Whether in computing means of rates and ratios we consciously use weights, or whether we do not, the resulting mean is weighted. The problem then becomes one of

devising a rational system of weights instead of leaving the result to chance weighting.[9]

The advantages and limitations of the arithmetic mean may now be summarized. It is (1) the most widely used of all averages; (2) it has a definite value; (3) it lends itself to algebraic treatment; (4) it is easily computed from either ungrouped or grouped data; (5) unless some other form of an average is specifically indicated or only a rough approximation to the central tendency is required, the mean is the best average to use. One caution should be borne in mind: the mean is sensitive to extreme values in the series and may not be truly representative, in which case some other measure of central tendency should be used along with it.

5. THE GEOMETRIC MEAN

For a series of items the geometric mean is the *n*th root of the product of the items. If the geometric mean is wanted for 10 items, the items are multiplied together and the 10th root taken. In terms of the formula, it may be expressed thus:

$$M_g = \sqrt[n]{(x_1)\,(x_2)\,(x_3)\,\ldots\ldots\ldots\,(x_n)}$$

To take a simple example:

$$\begin{aligned} M_g &= \sqrt[3]{(3)\,(6)\,(9)} \\ &= \sqrt[3]{162} \\ &= 5.45 \end{aligned}$$

If there are many items and large numbers are involved, the difficulty in extracting the *n*th root becomes very great. In such cases logarithms may be used. The arithmetic mean of the sum of the logarithms of the items is the logarithm of the geometric mean of the items; the logarithms may be found by consulting a logarithmic table. The geometric mean may be computed for either ungrouped or grouped data. The formula differs slightly from that given above and is as follows:

$$\text{Log } M_g = \frac{\Sigma f \log m}{N}$$

In order to compare it with the arithmetic mean, the data from Table XXXIX will be used:

[9] For further discussion of weighting see Chaddock, *op. cit.*, pp. 193-196; Secrist, Horace, *op. cit.*, pp. 241-246; Yule, *op. cit.*, pp. 220-225.

TABLE XLII

THE GEOMETRIC MEAN AGE OF UNEMPLOYED WORKERS IN BOSTON COMPUTED WITH THE USE OF LOGARITHMS

Age in Years (1)	Mid-Point m (2)	Number f (3)	$\log m$ (4)	$f \log m$ (5)
10–14	12.5	13	1.096910	14.259830
15–19	17.5	1,745	1.243038	2169.101310
20–24	22.5	2,968	1.352183	4013.279144
25–29	27.5	2,448	1.439333	3523.487184
30–34	32.5	2,176	1.511883	3289.857408
35–39	37.5	2,323	1.574031	3656.474013
40–44	42.5	2,234	1.628389	3637.821026
45–49	47.5	2,195	1.676694	3680.343330
50–54	52.5	1,786	1.720159	3072.203974
55–59	57.5	1,544	1.759668	2716.927392
60–64	62.5	1,107	1.795880	1988.039160
65–69	67.5	723	1.829304	1322.586792
		21,262		33084.380563

$$
\begin{aligned}
\text{Log } M_g &= \frac{33084.380563}{21262} \\
&= 1.556029 \\
&= 35.97 \text{ years}
\end{aligned}
$$

The geometric mean age is smaller by 2.7 years than the arithmetic mean. It is characteristic of the geometric mean that it gives less weight to extreme deviations than does the arithmetic mean, which results in a somewhat lower average. In the above problem it will be noticed that the logarithm of the mid-point of each class-interval is taken. Then the logarithm of this number is multiplied by the frequency of the class-interval. The sum of these products divided by the total frequencies gives the logarithm of the geometric mean.

Some social series show an aggregate increase over a period of time. Such are population, per capita income in the United States, and publicly supported social welfare activities. If it is assumed that the rate of change is the same in each year of a period under consideration and this rate of change is unknown but is to be determined, then the geometric method is the one to apply. The formula is as follows:

P_1 = population at the end of n years
P_0 = population at beginning of period
r = rate of change per year

n = number of years, used as the power to which the expression in the parentheses is to be raised

Or, using logarithms, the formula may be written:

$$\log P_1 = \log P_0 + n \log(1 + r), \text{ amount of change}$$

$$\log(1 + r) = \frac{\log P_1 - \log P_0}{n}, \text{ rate of change}$$

Where a power larger than a cube is used, the student will find the use of logarithms indispensable. Suppose we want to know the annual rate of growth of the population of the United States from 1920 to 1930. The substitutions would be as follows:

$$\begin{aligned}
\log(1 + r) &= \frac{\log(122{,}775{,}046) - \log(105{,}710{,}620)}{10.25} \\
&= \frac{8.089092 - 8.024116}{10.25} \\
&= .006339, \text{ logarithm of the rate} \\
(1 + r) &= 1.0147 \\
r &= 1.0147 - 1 \\
&= .0147, \text{ or } 1.47 \text{ per cent increase per year}
\end{aligned}$$

The importance of the geometric mean in estimating this type of change is emphasized by the fact that the arithmetic mean would be 1.57 per cent for the period of 10.25 years. The geometric mean allows for the changing volume of population each year, while the arithmetic mean uses the population of 1920 as 100 per cent for each succeeding year.

The investigator should be cautioned regarding this use of the geometric mean, however. As a matter of fact, population does not change at the same rate each year over a long period of time. Its growth is affected by immigration laws, by the spread of birth control, by the business cycle, and by wars. Other social series which show an upward trend over a long period of time also may have irregular rates of change. Consequently, the choice of the geometric mean as a single method of estimating the rate of change depends upon the judgment of the worker as to whether or not it really is the best method. Yule points out that, even if the geometric mean rate of change in population be a close approximation to the facts for a whole country, it cannot be assumed to represent the rate of change in smaller geographic subdivisions; these have special conditions which affect their rates of change.

The worker must constantly exercise his judgment to avoid unwarranted assumptions.[10]

The geometric mean, then, has some uses, in which it is superior to other averages. It can be used for estimating change in an augmenting social series, and it is useful in averaging ratios such as index numbers. It has the disadvantage, however, of being unfamiliar to many users of statistics and for that reason should be used with caution and with full explanation of its significance.

6. RELATIONS EXISTING AMONG AVERAGES

The quantitative relations existing among the four averages discussed above are not constant, but in some types of frequency distributions they approximate to constant conditions. The relations of the mean, median, and mode are determined by the degree of asymmetry of the frequency distribution.

It has been noticed by Pearson and others that in certain moderately asymmetrical distributions the median is located at a point between the mean and the mode about one-third the distance from the mean in the direction of the mode, and the rule has been laid down that this may be taken as a fairly constant relation among the three measures of central tendency. In view of this fact, Pearson has proposed the following formula for determining a rough measure of the mode in moderately asymmetrical distributions:

$$Mo = M - 3(M - Md)$$

Obviously the mode computed by this formula will be twice as far from the median as the median is from the mean. The question to be raised regarding any frequency distribution is whether or not it is "moderately asymmetrical." Preceded by the word "moderately," this concept becomes qualitative and not quantitative. If defined in quantitative terms, it should mean any distribution having a mode, determined by more exact methods than the method under discussion, twice as far from the median as the median is from the mean. The distribution presented in Figure L appears to the eye to be moderately asymmetrical but, when defined in terms of the above mean-median-mode relation, it is clear that it is not moderately asymmetrical because the median and the mode are very close together, while both are considerably higher than the mean. The important point concerning the relative

[10] Yule, *op. cit.*, p. 126.

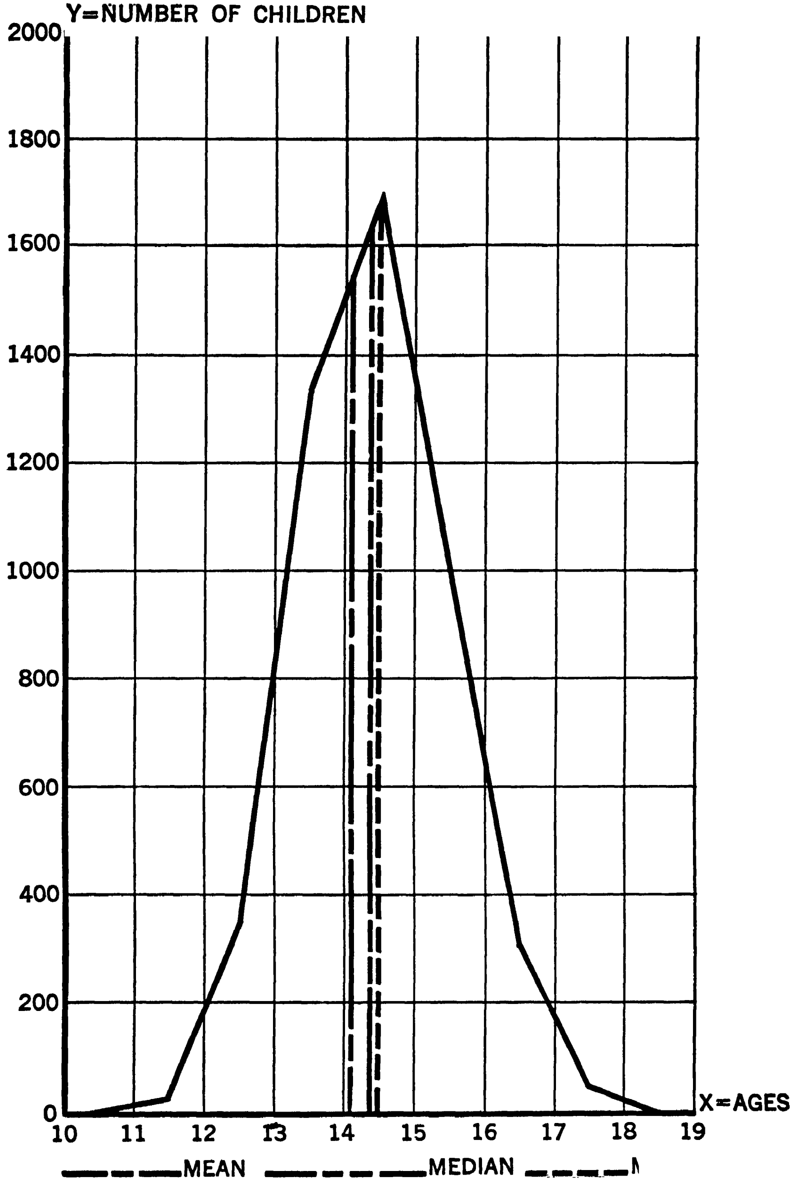


FIGURE L.—DISTRIBUTION OF CHILDREN IN THE EIGHTH GRADE, ST. LOUIS PUBLIC SCHOOLS, BY AGES: GRAPHIC LOCATION OF THE MEAN, MEDIAN, AND MODE

positions of the mean, median, and mode is that the median and the mode always move in the direction of the skew of the frequency distribution. Consequently, they may be used along with the mean and the standard deviation from the mean as measures of skewness.

To illustrate the relative positions of the mean, median, and mode in a moderately asymmetrical distribution the data, from Table XIX are presented in graphic form above with the three averages indicated.

For the data presented in Figure L the mean is 14.05 years, the median 14.33 years, and the mode 14.42 years. The three measures of central tendency are very close together, because this frequency distribution is only moderately asymmetrical.

There is no constant relation between the mean and the geometric mean, except that the geometric mean is always somewhat smaller. This is due to the fact that in squaring the quantities to obtain the geometric mean the extreme values are minimized. The degree of difference between the mean and the geometric mean will vary directly with the ratio of the standard deviation to the mean.[11]

7. EXERCISES

1. Data for computation of averages:

TABLE XLIII

DISTRIBUTION BY AGES OF PAROLEES, CLASSIFIED BY TOTAL AND BY SUCCESS[1]

Age in Years (1)	Parolees, All Classes (2)	Parolees Successful at Each Age (3)
Total	1,004	557
7	1	1
8	1	1
9	4	1
10	8	4
11	16	7
12	28	15
13	41	21
14	87	43
15	111	65
16	190	100
17	190	102
18	126	69

[11] Yule, *op. cit.*, pp. 123, 156.

TABLE XLIII—(*Continued*)

Age in Years (1)	Parolees, All Classes (2)	Parolees Successful at Each Age (3)
19	96	56
20	44	31
21	23	14
22	15	11
23	7	5
24	5	4
25	6	5
26	1	1
27	1	1
28	2	0
29	0	0
30	1	0

[1] *Missouri Crime Survey*, p. 469. New York: Macmillan, 1926. Table XLIII is derived from Table XVI of this report.

(a) Find the modal age for all parolees and for successful parolees by a graphic method and by a formula. Compare the modal age in the two groups.
(b) Find the median age for each group in the table using the cumulative frequency curve method and a formula. Compare the median ages found.
(c) Find the mean age of parolees in each column of the table, using both the long and the short method. Compare the two means. Why are they the same or approximately the same size? Could you call this a weighted mean?
(d) Compare the mode, median, and mean found for each column of data.
(e) Do these ages of parolees illustrate a symmetrical, moderately asymmetrical or highly asymmetrical distribution? How would you determine this, using only statistical methods thus far described?

Data for computation of averages:
(a) Compute the mode, median, mean, and geometric mean wage for the 423 wage earners in this table.
(b) By adding pairs of class-intervals, increase the class-interval to $200. Compute the mode, median, mean, and geometric mean wage from the results and compare the averages with (a).
(c) Would you call this a moderately asymmetrical frequency distribution?

TABLE XLIV

EARNINGS OF CHIEF WAGE EARNERS IN FAMILIES [1]

Earnings of Wage Earner	Number of Wage Earners
T	423
$ 800– 899	6
900– 999	11
1,000–1,099	40
1,100–1,199	50
1,200–1,299	63
1,300–1,399	63
1,400–1,499	81
1,500–1,599	45
1,600–1,699	24
1,700–1,799	20
1,800–1,899	6
1,900–1,999	7
2,000–2,099	2
2,100–2,199	4
2,200–2,299	0
2,300–2,399	1

[1] Houghteling, Leila, *The Income and Standard of Living of Unskilled Laborers in Chicago*, p. 27. University of Chicago Press, 1927.

8. REFERENCES

Chaddock, Robert E., *Principles and Methods of Statistics*, Chaps. VI, VII.

Kelley, Truman L., *Statistical Method*, Chap. III.

Mills, Frederick C., *Statistical Methods*, Chap. IV.

Secrist, Horace, *An Introduction to Statistical Methods*, Chap. IX.

Yule, G. U., *An Introduction to the Theory of Statistics*, Chap. VII.

CHAPTER IX

Measures of Dispersion

1. INTRODUCTION

IN THE preceding chapter we have been concerned with the tendency of values in social data to cluster around a central value. Measures of this tendency are useful in arriving at a shorthand description of the data. But the tendency of data to scatter below and above the central value is as noticeable as is concentration. An adequate description of a frequency distribution requires knowledge of both scatter and concentration. Scatter is usually referred to in statistics as dispersion or variation. Deviations from the central value may be due to chance; that is, the whole universe of a particular type of data, if it could be taken into consideration, would show dispersion about the average. Deviations from an average may be due to the method by which the sample was selected from the universe of similar data. The sample may not fairly represent the universe from which it was selected, in which case the amount of dispersion may be either less or more than it would be for the universe. Thus, deviations from the central value are due both to chance and to the method of sampling.

Measures of dispersion are practical checks on the homogeneity of the data. The smaller the amount of dispersion around the average, the greater the homogeneity of the data for the trait measured. Conclusions drawn from the study of relatively homogeneous data are more reliable than those drawn from the study of data which are highly heterogeneous. The amount of dispersion for a given sample may be less than the amount found in an absolutely random sample from the universe of the same kind of social phenomena. The measure of dispersion shows this fact, but at the same time it indicates a high degree of homogeneity, and conclusions drawn for this sample but not extended to any other data of the same universe will be correspondingly reliable. This may

be illustrated in various ways. For a number of years effort has been made by psychologists to find empirically a normal distribution of intelligence among a sample of children. Terman found that a sample of 905 intelligence quotients, which he and his associates obtained, was distributed approximately as a bell-shaped curve. The measure of the dispersion of this random sample might then be taken as a close approximation to the measure of dispersion of intelligence quotients, if all children in the United States were examined. Certain school policies might be based upon the dispersion of intelligence in this random sample, but the amount of dispersion would be greater than it would be for a sample of children attending a school which selects only children with I.Q's, say, at or above 110; and it would be higher than the amount of dispersion found among children in a school for the feeble-minded. Conclusions based upon the amount of dispersion of the I.Q's would be more reliable in the formulation of specific policies for these schools than would conclusions affecting the policies of a school whose children had a normal distribution of I.Q's. The homogeneity of intelligence among the children of the two schools would be high. The dispersion found in the age distribution of workers in an industry is a measure indicating the policy of the industry to restrict employment to certain age groups or to disregard age. Compared with dispersion of ages in the general population, the dispersion in a particular industry might be small; this would suggest, as a matter for further study, that possibly there is discrimination against workers over a certain age limit.

A measure of dispersion may assist public health officials to judge the effectiveness of their work. A city which has the census tract system and uses the census tracts as public health units will serve as an illustration of this use of measures of dispersion. An average death rate for all tracts may be computed and the dispersion found. Those tracts which deviate widely from the average rate have either exceptionally good or exceptionally bad health conditions. Those in which mortality is high, assuming a standard population has been used for computing rates, must have some health disadvantages. The location of these tracts by means of their dispersion from the average rate enables the health officials to concentrate efforts at those points which need improvement most. Thus measures of dispersion become aids to social control.

One other use of measures of dispersion may be mentioned. In all measures of the degree of interrelationship between sets of social phenomena some measure of dispersion has to be used, because relationship is expressed as a function of average variability, involving both direction and amount of variability. This use of measures of dispersion will become apparent when we take up the subject of correlation.

2. THE QUARTILE DEVIATION

The first and third quartiles of a frequency distribution indicate dispersion from the median as the average. The first quartile is the value of the item below which 25 per cent of the values fall, and the third quartile is the value of the item above which 25 per cent of the items fall. That is, between the first and third quartiles half the items in the frequency distribution are found. Like the median, the quartiles are position values. In order to determine their values the data must be arranged in class-intervals from lowest to highest values. The quartiles are not averages; they do not represent central tendency. They represent deviations from central tendency. For that reason they properly belong under the discussion of measures of dispersion. The quartile deviation is the sum of the first and third quartiles divided by 2. The values between the first and third quartiles are sometimes referred to as the interquartile range, and the quartile deviation as the semi-interquartile range.

Before the quartile deviation can be determined, the values of the first and third quartiles must be computed. They may be found by formulas similar to that used for locating the median (p. 234):

$$f$$

$\mathcal{Q}_1$ = first quartile
l = lower limit of the class-interval in which the first quartile falls
N = total number of items in the frequency distribution
F = sum of all frequencies in classes below l
i = value of the class-interval
f = number of items in the class-interval containing the first quartile

Using the data for unemployed men in Boston and referring to Table XXXVII, the substitutions would be as follows:

$$2448$$

$$= 26.2 \text{ years}$$

Q_1 is 26.2 years. That is, 25 per cent of the unemployed workers in Boston were 26.2 years of age or less. The formula for determining the third quartile may be written as follows:

$$\mathcal{Q}_3 = l + \frac{\frac{\quad}{4} - F}{f}\, i$$

In this formula the meaning of the symbols is not changed except that l refers to the lower limit of the class-interval in which the third quartile falls. The other symbols may be read as in the preceding formula. The only other change is in the multiplication of n by 3 in order to obtain three-fourths of the total frequencies, reading upward from the lowest toward the highest. Substituting in this formula, we get:

$$\mathcal{Q}_3 = 45 + \frac{\frac{\quad}{4} - 13907}{2195}\, 5$$

$$= 45 + \frac{15946 - 13907}{2195}\, 5$$

$$= 49.6 \text{ years}$$

Q_3 is 49.6 years. Seventy-five per cent of the unemployed workers in Boston were 49.6 years of age or less.

The formula for the quartile deviation is:

$$\mathcal{Q} = \frac{\mathcal{Q}_3 - \mathcal{Q}_1}{2}$$

Substituting the values of the first and third quartiles found for the unemployment data in this formula, we get:

$$\mathcal{Q} = \frac{49.6 - 26.2}{2}$$

$$= 11.7 \text{ years, the quartile deviation}$$

If the data are ungrouped and are arranged in an array, the first and third quartiles are easily determined by simply counting

off from the lowest value 25 and 75 per cent of the items, respectively. The formula for the quartile deviation may then be used.

The advantages of the quartile deviation as a measure of dispersion are that it is a definite quantity, easily computed, and simple to understand. But it is a position measure of dispersion and does not lend itself to algebraic uses. Another limitation of the quartile deviation is the fact that it is not affected by the variability of the items whose values lie either between the first and third quartiles or outside of them. The quartile deviation is simply the mean deviation of the values of the first and third quartiles. If for special reasons the median is preferred as the average to be used, then the logical measure of dispersion to use with it is the quartile deviation. Otherwise, it is probably better to employ some other measure of dispersion.

3. PERCENTILES AND DECILES

Another measure of dispersion which resembles the quartile deviation in being a position value is the percentile. A percentile is a rank on a scale divided into 100 equal parts, and the value of any particular percentile is equal to the sum of the hundredths below and including the particular rank. It is a percentage concept. Deciles are simply the 10th, 20th, 30th, etc., percentiles. If a position measure of dispersion is to be used, percentiles or deciles are in some respects preferable to the quartiles, because they give a more detailed description of dispersion. For certain technical purposes the percentile measure of dispersion has been found very useful. Perhaps it has been used most by psychologists and educational administrators for ranking school children according to intelligence or school ability. Some psychologists prefer to rank the children tested on a percentile scale rather than to assign I.Q's. The percentile method may also be used for such purposes as ranking rates of piece workers in a factory, death rates by counties, birth rates by counties, crime rates by census tracts, etc. There is no statistical reason why the percentile method could not be applied to any type of data, but in practice its use has been confined largely to educational and psychological data. Yule suggests that it may also be used to show the distribution of non-measurable traits.[1]

The computation of percentiles will be illustrated from the following table which gives the infant mortality rates in 1929 for 108 cities of the United States:

[1] Yule, *op. cit.*, p. 150.

TABLE XLV

PERCENTILE DISTRIBUTION OF INFANT MORTALITY RATES IN 108 CITIES IN THE UNITED STATES, 1929[1]

Infant Death Rate	f	f Cumulated	Percentiles at Mid-Point of Class-Interval
(1)	(2)	(3)	(4)
30– 34.9	1	1	0.47
35– 39.9	2	3	1.86
40– 44.9	1	4	3.27
45– 49.9	9	13	8.41
50– 54.9	4	17	13.95
55– 59.9	17	34	23.77
60– 64.9	11	45	36.79
65– 69.9	20	65	51.15
70– 74.9	18	83	68.82
75– 79.9	4	87	79.05
80– 84.9	4	91	82.77
85– 89.9	1	92	85.10
90– 94.9	2	94	86.49
95– 99.9	3	97	88.82
100–104.9	0	97	90.21
105–109.9	1	98	90.78
110–114.9	2	100	92.12
115–119.9	2	102	93.93
120–124.9	2	104	95.79
125–129.9	1	105	97.19
130–134.9	1	106	98.17
135–139.9	2	108	99.51

[1] *Weekly Health Index*, United States Bureau of the Census, Vol. II, No. 35.

The formula for computing any percentile is:

$$f$$

P = the value of the percentile to be found
l = value of lower limit of class in which percentile occurs
p = Per cent of cases having values equal to or less than P
N = number of items in entire frequency distribution
F = total frequencies below particular percentile class-interval
f = frequencies in particular percentile class-interval
i = value of the class-interval

The similarity between this formula and the formula for the median is apparent. In each case the aim is to determine the value of an item at a certain position in a frequency distribution. In column (4) of Table XLV the percentiles which fall approximately at the middle of the class-intervals are given; the infant death rate is, of course, the mid-point of the class-interval opposite the percentile concerned. But how could the value of the 40th percentile be determined? The first thing to determine is the

class-interval in which the 40th percentile falls. Forty per cent of the 108 rates will be below the 40th percentile, and 40 per cent of 108 is 43.2. Hence, the 40th percentile falls in the class-interval 60-64.9 because there are 45 frequencies below 64.9. Now, we may substitute in the formula:

$$P = 60 + \frac{(40.0)\,(108) - 34}{11}\,5$$

11

$$= 64.2, \text{ the 40th percentile death rate}$$

Any other percentile may be found in like manner.

If a percentile curve is constructed for a set of data, any percentile may be located graphically with a fair degree of accuracy. The form of the percentile curve is an ogive, such as that below in Figure LI. The percentiles for the mid-points of the class-intervals in Table XLV were used to plot this curve.

The broken horizontal and vertical lines on the graph were drawn to locate the value of the 40th percentile. A line was drawn from the zero ordinate along the 40th abscissa until it intersected the curve. At this point a perpendicular line was dropped to the base. This perpendicular intersects the base line slightly above the 64th ordinate. That is, the value of the 40th percentile is a little more than 64—by formula it was found to be 64.2.

Sometimes the only percentiles wanted are the deciles, or every tenth percentile. A decile is determined in the same manner as any other percentile.

The principal value of percentiles and deciles as measures of dispersion lies in their simplicity. We are accustomed to think in terms of percentage and tenths. Consequently, when it is said that 40 per cent of the cities reporting infant mortality to the Bureau of the Census have rates of 64.2 or less, little explanation is required. That is essentially what the 40th percentile means. If the 90th percentile has a value of 100, we know that 10 per cent of the cities had infant mortality rates greater than 100, which is very high. To give the values of the deciles or the values of several percentiles at other points is to suggest the degree of dispersion below and above the median percentile. In the case of intelligence ratings, the use of percentiles instead of I.Q's may reflect a healthy skepticism of intelligence tests and convey the meaning that the examiner is discussing only the distribution of intelligence in the group examined and that he is distributing

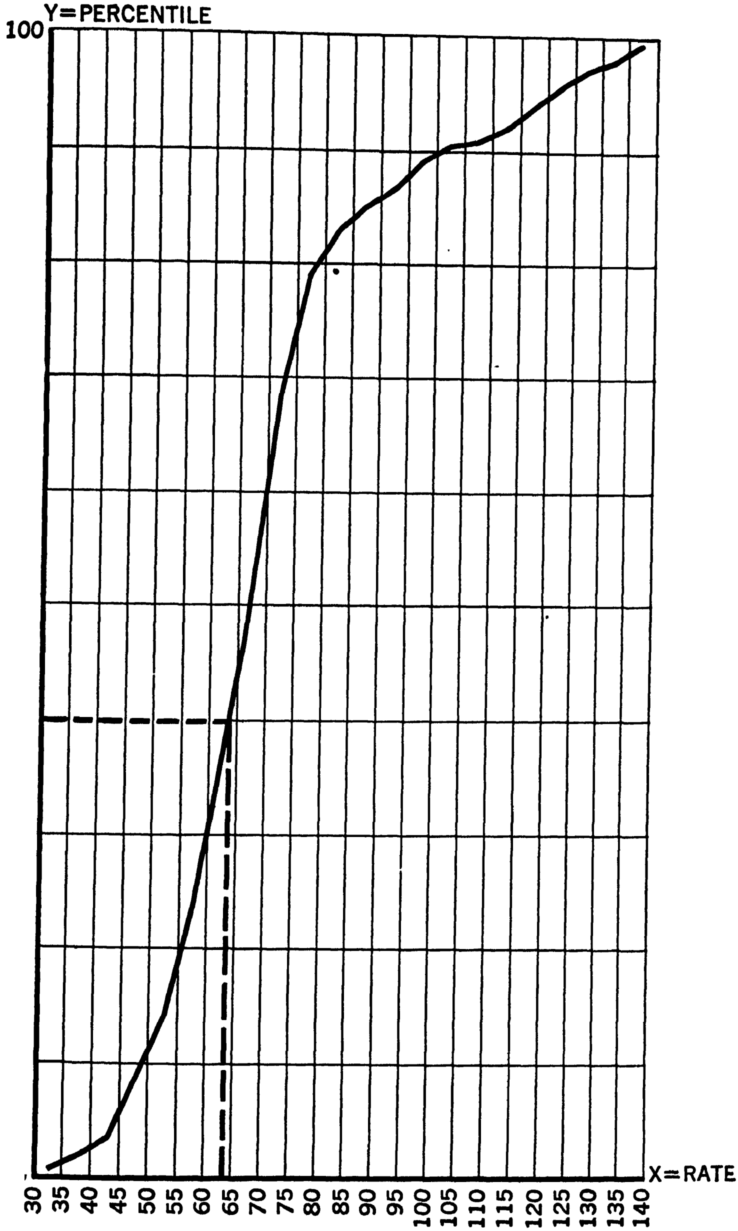


FIGURE LI.—PERCENTILE DISTRIBUTION OF INFANT MORTALITY RATES IN 108 CITIES IN THE UNITED STATES, 1929

ratings of what the tests test, whether it be intelligence or something else.

4. THE AVERAGE, OR MEAN, DEVIATION

The average deviation is the mean of the deviations from an average, disregarding algebraic signs. It may be computed from the mean, median, or mode, but generally the mean or the median is used. The sum of the deviations from the median is slightly less than the sum of the deviations from the mean. Hence, the average deviation is somewhat smaller when computed from the median than when computed from the mean. For this reason many statisticians think it best to use the median from which to compute the average deviation, unless practical considerations make the mean the preferable average.[2] Both methods will be illustrated.

The average deviation may be computed from either ungrouped or grouped data. Computation from ungrouped data will be illustrated from the amount of relief per case given by twenty family relief agencies in cities reporting to the Russell Sage Foundation:

TABLE XLVI

COMPUTATION OF THE AVERAGE DEVIATION FROM UNGROUPED DATA: AMOUNTS OF RELIEF PER RELIEF CASE IN 20 FAMILY RELIEF AGENCIES IN JULY, 1931

Relief per Relief Case X	Deviations from Mean d	Deviations from Median d
\$26.85	−\$ 3.196	\$ .615
28.99	− 1.056	2.755
48.62	18.574	22.385
34.31	4.264	8.075
24.30	− 5.746	− 1.935
12.31	− 17.736	− 13.925
34.40	4.354	8.165
25.62	− 4.426	− .615
45.92	15.874	19.685
38.59	8.544	12.355
52.45	22.404	26.215
24.05	− 5.996	− 2.185
21.17	− 8.876	− 5.065
24.23	− 5.816	− 2.005
36.64	6.594	10.405
19.08	− 10.966	− 6.155
18.61	− 11.436	− 7.625
42.99	12.944	16.755
17.78	− 12.266	− 8.455
24.01	− 6.036	− 2.225
	\$187.104	\$177.600
	M = 30.046	Md = 26.235

[2] Yule, *op. cit.*, p. 145.

The algebraic signs have been inserted to show that the algebraic sum of the deviations from the mean is zero—the deviations are actually -93.552 and $+93.552$. But the algebraic sum of the deviations from the median is not zero—the deviations are -50.190 and $+127.410$. Since signs are neglected in computing the average deviation, the inequality of the plus and minus deviations from the median does not affect the result. It should be noted that it is necessary to carry the deviations from the mean to three decimal places in order to make the plus and minus deviations equal.

The formula for the computation of the average deviation from either mean or median is:

A. D. = average deviation
d = deviation from the average
N = number of items

Substituting in this formula the values for deviations from the mean, we have:

$$\text{A. D.} = \frac{187.104}{20}$$
$$= 9.355$$

Using the values of the deviations from the median, we have:

$$\text{A. D.} = \frac{177.600}{20}$$
$$= 8.862$$

The average deviation from the mean is .493 larger than the average deviation from the median. This indicates that the values of the items cluster a little more closely about the median than they do about the mean. But the location of the median precludes the full influence of the higher deviations, as can be seen from the excess of plus over minus deviations from the median. If the purpose of the worker is to allow full weight to all deviations, then the average deviation from the mean would be the one to use. If he wants to emphasize the value from which the sum of the deviations is least, then he should use the median.

It does not often happen in practice that the data used are ungrouped. For that reason it is necessary to have a method for

computing the average deviation from grouped data. The long method of computing the average deviation will be illustrated first. The data used will be the ages of unemployed workers in Boston. Table XLVII shows the details of this method:

TABLE XLVII

COMPUTATION OF THE AVERAGE DEVIATION FROM THE MEAN AND FROM THE MEDIAN FOR THE AGES OF UNEMPLOYED WORKERS IN BOSTON

Age (1)	m (2)	Deviations from Mean— $(m - M)$ d (3)	f (4)	fd (5)	From Median $(m - Md)$ d (6)	fd (7)
10–14	12.5	26.2	13	340.6	25.3	328.9
15–19	17.5	21.2	1,745	36,994.0	20.3	35,423.5
20–24	22.5	16.2	2,968	48,081.6	15.3	45,410.4
25–29	27.5	11.2	2,448	27,417.6	10.3	25,214.4
30–34	32.5	6.2	2,176	13,491.2	5.3	11,532.8
35–39	37.5	1.2	2,323	2,787.6	.3	696.9
40–44	42.5	3.8	2,234	8,489.2	4.7	10,499.8
45–49	47 5	8.8	2,195	19,316.0	9.7	21,291.5
50–54	52.5	13.8	1,786	24,646.8	14.7	26,254.2
55–59	57.5	18.8	1,544	29,027.2	19.7	30,416.8
60–64	62.5	23.8	1,107	26,346.6	24.7	27,342.9
65–69	67.5	28.8	723	20,822.4	29.7	21,473.1
			21,262	257,760.8		255,885.2
	$M = 38.7$				$Md = 37.8$	

The principal difference in the computation from grouped data as compared with ungrouped data is that the deviations from the average are taken from the mid-value of the class-interval and then multiplied by the class frequencies. The deviations from *m*, the mid-values, are shown in columns (3) and (6), and the frequencies are given in column (4). The products of the deviations and the frequencies are given in columns (5) and (7). The results are:

$$\text{A. D.} = \frac{\Sigma fd}{N}$$

Using the mean: $\text{A. D.} =$
$$21{,}262$$
$$= 12.1 \text{ years}$$

Using the median: $$\text{A. D.} = \frac{255{,}885.2}{21{,}262}$$
$$= 12.0 \text{ years}$$

There is in the two average deviations a difference of .1 of a year. This is so small as to be unimportant except for theoretical purposes.

This long method requires the use of large numbers and much labor. The same results can be obtained by using a short method for computing the average deviation. This short method is illustrated below:

TABLE XLVIII

COMPUTATION OF THE AVERAGE DEVIATION FOR THE SAME DATA BY THE SHORT METHOD

Age (1)	m (2)	f (3)		Step-Deviations from Assumed Mean d (4)	fd (5)	Step-Deviations from Assumed Median d (6)	fd (7)
10–14	12.5	13		−5	65	−5	65
15–19	17.5	1,745		−4	6,980	−4	6,980
20–24	22.5	2,968	11673	−3	8,904	−3	8,904
25–29	27.5	2,448		−2	4,896	−2	4,896
30–34	32.5	2,176		−1	2,176	−1	2,176
35–39	37.5	2,323		0		0	
40–44	42.5	2,234		1	2,234	1	2,234
45–49	47.5	2,195		2	4,390	2	4,390
50–54	52.5	1,786	9589	3	5,358	3	5,358
55–59	57.5	1,544		4	6,176	4	6,176
60–64	62.5	1,107		5	5,535	5	5,535
65–69	67.5	723		6	4,338	6	4,338
		21,262					51,052

The data above may be substituted in the following formula:

in which n_1 is the number of items for which deviations measured from the assumed average are smaller than deviations measured from the true average; n_2 is the number of items for which deviations measured from the assumed average are larger than deviations measured from the true average; c is the difference between the assumed average and the true average; and i is the value of the class-interval. Since the deviations above are not in terms of years but in terms of steps, or class-intervals, c must be expressed as a fraction of a step:

$$\text{A. D.} = \frac{51{,}052 + (11{,}673 -}{21{,}262}$$

$$= 12.1 \text{ years (using the mean)}$$

$$\text{A. D.} = \frac{51{,}052 + (11{,}673 - 9{,}589).06}{21{,}262} \; 5$$

$$= 12.1 \text{ years (using the median)}$$

The results by the short method are identical with those obtained by the long method. In the illustration the small exceed the large deviations from the average because the true mean, 38.7, is nearer the upper limit of class 35-39, than is the assumed mean, 37.5; but sometimes the situation will be reversed, in which case n_2 will be larger than n_1, and it will be necessary to subtract the correction factor instead of adding it. But this should be clear from the formula. If the expression inside the parentheses is a minus quantity, then the plus sign in front of the parenthesis leaves it a minus and indicates subtraction, because a plus times a minus gives a minus quantity. The deviations on the side of the assumed average will always be smaller than they should be. In the illustration the assumed average is less than the true average. Hence, all the frequencies in the class-interval containing the assumed average and all those in lower class-intervals will be too small by the amount of the correction factor. The deviations in all class-intervals higher than that in which the assumed average falls will be too large by the amount of the correction factor. Suppose the assumed average is higher than the true average. The rule still holds, but the small deviations are now at the upper end of the distribution, and the large deviations are at the lower end. (The average deviation from the mean and the median is the same to one decimal place in this problem, but this would not generally be true.)

Occasionally it may be desirable to obtain the average deviation of death rates in a city for a period of twenty years. The average deviation can be found by using the method for ungrouped data. A caution should be mentioned, however. Time series are complex. They generally show four types of variation: trend, cycle, seasonal fluctuation, and residual fluctuation. A measure of dispersion applied to time series usually means less than when applied to frequency distributions.

The average deviation is simple to compute. It may be computed from either grouped or ungrouped data, and either the mean or the median may be used as the average. Although useful, the average deviation is not employed as much, as a step in further statistical analysis, as is the standard deviation.

5. STANDARD DEVIATION

The standard deviation is the square root of the mean of the squares of the deviations from the arithmetic mean. It is never computed from any average but the mean. The concept of the standard deviation developed in connection with studies of the normal curve of error during the nineteenth century. Efforts to measure the probability of error due to chance resulted in the concepts known as the "modulus," the "mean error," and the "probable error." Working with biological data, Karl Pearson found it more convenient to work with the concept to which he gave the term, standard deviation.[3] The method had been used before this time, but Pearson's use has given it currency. The standard deviation enters into so much statistical analysis that it is particularly important for the student to understand its meaning and its method of computation.

The method of computation is similar to that of the average deviation, except that the deviations are squared, which disposes of the algebraic signs by making all signs plus. The standard deviation may be computed from grouped or ungrouped data, and it may be computed by both the short and the long method. The long method will be illustrated first by the use of the ages of unemployed workers in Boston.

TABLE XLIX

COMPUTATION OF THE STANDARD DEVIATION OF THE AGES OF UNEMPLOYED WORKERS IN BOSTON BY THE LONG METHOD

Age (1)	m (2)	f (3)	d (4)	d^2 (5)	fd^2 (6)
10–14	12.5	13	26.2	686.44	8,923.72
15–19	17.5	1,745	21.2	449.44	784,272.80
20–24	22.5	2,968	16.2	262.44	778,921.92
25–29	27.5	2,448	11.2	125.44	307,077.12
30–34	32.5	2,176	6.2	38.44	83,645.44
35–39	37.5	2,323	1.2	1.44	3,345.12
40–44	42.5	2,234	3.8	14.44	32,258.96
45–49	47.5	2,195	8.8	77.44	169,980.80
50–54	52.5	1,786	13.8	190.44	340,125.84
55–59	57.5	1,544	18.8	353.44	545,711.36
60–64	62.5	1,107	23.8	566.44	627,049.08
65–69	67.5	723	28.8	829.44	599,685.12
		21,262			4,280,997.28

[3] Walker, Helen M., *Studies in the History of Statistical Method*, pp. 52-54, 64. Williams & Wilkins, Baltimore, 1929.

The symbols in this table have the same meaning which they have in the formula for the average deviation, and the general formula for the standard deviation computed from grouped data is as follows:

$$\sqrt{\frac{\Sigma fd^2}{N}}$$

Small sigma is the symbol for the standard deviation. Substituting the data from Table XLIX in this formula, we have:

21,262

= 14.2

The standard deviation is somewhat larger than the average and the quartile deviations. The relations of these three measures of dispersion will be discussed later in the chapter.

If the data are ungrouped, the procedure is simple. The formula is

in which d is the deviation from the arithmetic mean. The sum of the squared deviations from the mean is divided by the number of items, and the square root of the result is taken, giving the standard deviation.

TABLE L

COMPUTATION OF THE STANDARD DEVIATION OF THE AGES OF UNEMPLOYED WORKERS IN BOSTON BY THE SHORT METHOD

Age (1)	m (2)	f (3)	Steps d (4)	$-fd$ (5)	$+fd$ (6)	fd^2 (7)
10–14	12.5	13	−5	65		325
15–19	17.5	1,745	−4	6,980		27,920
20–24	22.5	2,968	−3	8,904		26,712
25–29	27.5	2,448	−2	4,896		9,792
30–34	32.5	2,176	−1	2,176		2,176
35–39	37.5	2,323	0			
40–44	42.5	2,234	1		2,234	2,234
45–49	47.5	2,195	2		4,390	8,780
50–54	52.5	1,786	3		5,358	16,074
55–59	57.5	1,544	4		6,176	24,704
60–64	62.5	1,107	5		5,535	27,675
65–69	67.5	723	6		4,338	26,028
		21,262		23,021	28,031	172,420

There is even more reason for using a short method of computing the standard deviation than for computing averages or the average deviation, because squaring the deviations from the mean increases the size of the numbers handled to very large quantities. The short method is illustrated in the next table.

$$= \sqrt{\frac{172{,}420}{21{,}262} - \left(\frac{28{,}031 - 23{,}021}{21{,}262}\right)^2}$$

$$= \quad 2.84 \text{ step deviations}$$

Multiplying by 5, $= \quad 14.2$ years

Using the short method, the standard deviation is identical with the standard deviation computed by the long method. But the numbers handled are smaller, and this makes for rapidity of computation and reduces the chances of error. It should be noted that the correction factor computed in the use of the short method is always subtracted from the sum of the fd^2's divided by the sum of the items, and that it is squared before subtracting. As suggested above, the reason for this is that the square of the deviations from the arithmetic mean is a minimum. It follows, therefore, that, if any correction is required, it must be because the sum of the deviations from the assumed mean is too large and, hence, must be decreased by the amount of the correction factor. In the above case the assumed mean is 37.5, whereas the true mean is 38.7. The result is that each deviation is too large by the amount of the correction factor. Since the deviations under the radical are already squared, it follows that the correction factor must be squared before deduction.

We may summarize the advantages which make the standard deviation preferable to any other measure of dispersion, unless special reasons exist for using some other measure. Squaring removes the differences of signs and gives weight to extreme variations. The standard deviation lends itself to algebraic treatment, is rigidly defined, is based upon all observations, is the most commonly used measure of dispersion, and is a step in many other statistical procedures. The squaring and extraction of the square root may appear to be rather complicated, but practice reduces this apparent difficulty, and the use of a table of squares and

[8a] The correction factor, c^2, is $\left(\frac{\Sigma fd}{N}\right)^2$.

square roots reduces the labor to a matter of listing the squares and roots.

6. RELATIONS OF Q, A.D., AND σ

In a perfectly symmetrical frequency distribution constant relations exist among the quartile, the average, and the standard deviation. It is rare in social statistics to find even a close approximation to a symmetrical distribution, but some distributions are sufficiently symmetrical to make significant comparisons with moderately asymmetrical distributions. The following table gives the ratios of each of the three measures of dispersion to the others, as computed by Thorndike:

TABLE LI

THE RELATIVE VALUES OF THREE MEASURES OF DISPERSION

Measures of Dispersion (1)	Perfectly Symmetrical Distribution[1] (2)	Ages of 21,262 Unemployed Workers (3)	Differences (2) − (3) (4)
σ	1.2533 times A. D.	1.1736 times A. D.	.0797
σ	1.4825 " Q	1.2137 " Q	.2688
A. D.	.7979 " σ	.8521 " σ	−.0542
A. D.	1.1843 " Q	1.0342 " Q	.1501
Q	.6745 " σ	.8239 " σ	−.1494
Q	.8453 " A. D.	.9669 " A. D.	−.1216

[1] Thorndike, E. L., *Mental and Social Measurements*, 2 ed., 1913, p. 67.

The differences between a perfectly symmetrical distribution and the distribution of ages of the unemployed workers are not large but they suggest a considerable variation of the latter from the bell-shaped curve. The ideal curve is a norm to which other curves approach more or less closely.

The differences between the different measures of dispersion are shown graphically below:

It is clear from the diagrams that plus and minus once the quartile deviation from the median, plus and minus once the average deviation from the mean, and plus and minus once the standard deviation from the mean include an increasing proportion of all the items in the order named. In a perfectly symmetrical distribution 50 per cent of all the items fall between the value equal to the median minus Q and the value equal to the median plus Q. In a perfectly symmetrical distribution 57.5 per cent of all the items are included between the value equal to the mean or median minus the average deviation and the value equal to the mean or

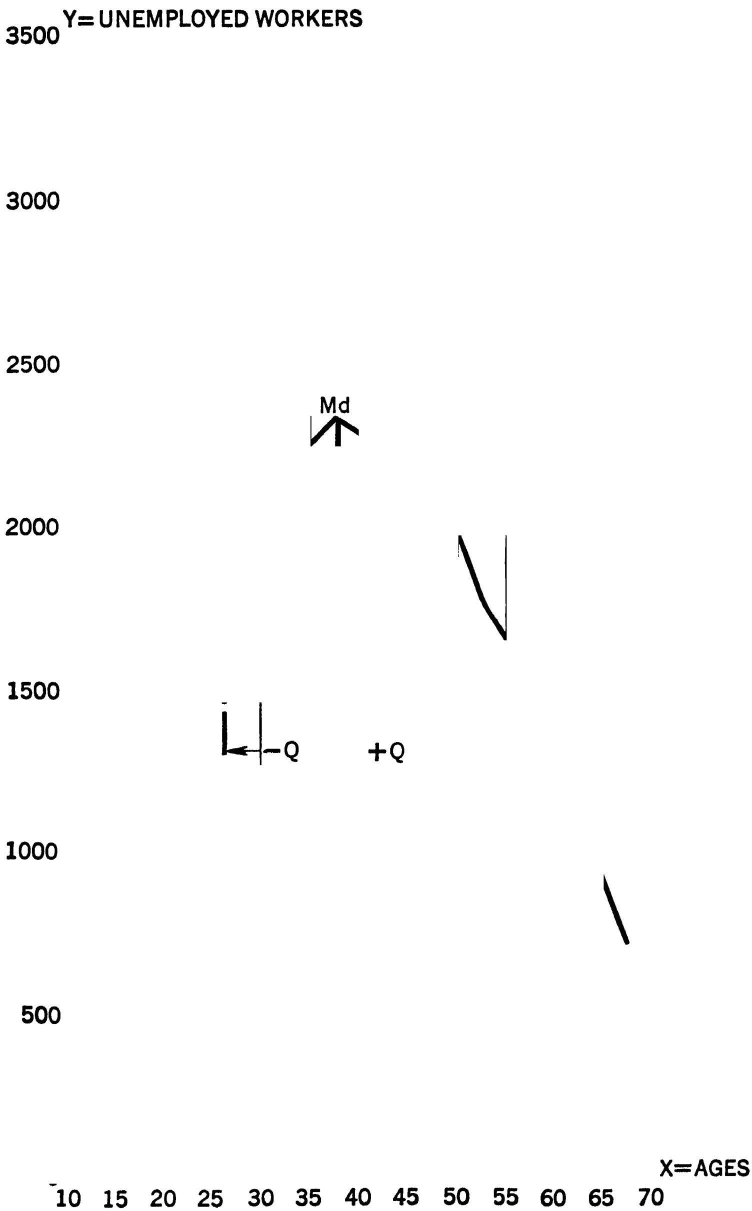


FIGURE LII.—AREA OF SURFACE ENCLOSED BY PLUS AND MINUS ONCE THE QUARTILE DEVIATION FROM THE MEDIAN AGE OF BOSTON WORKERS

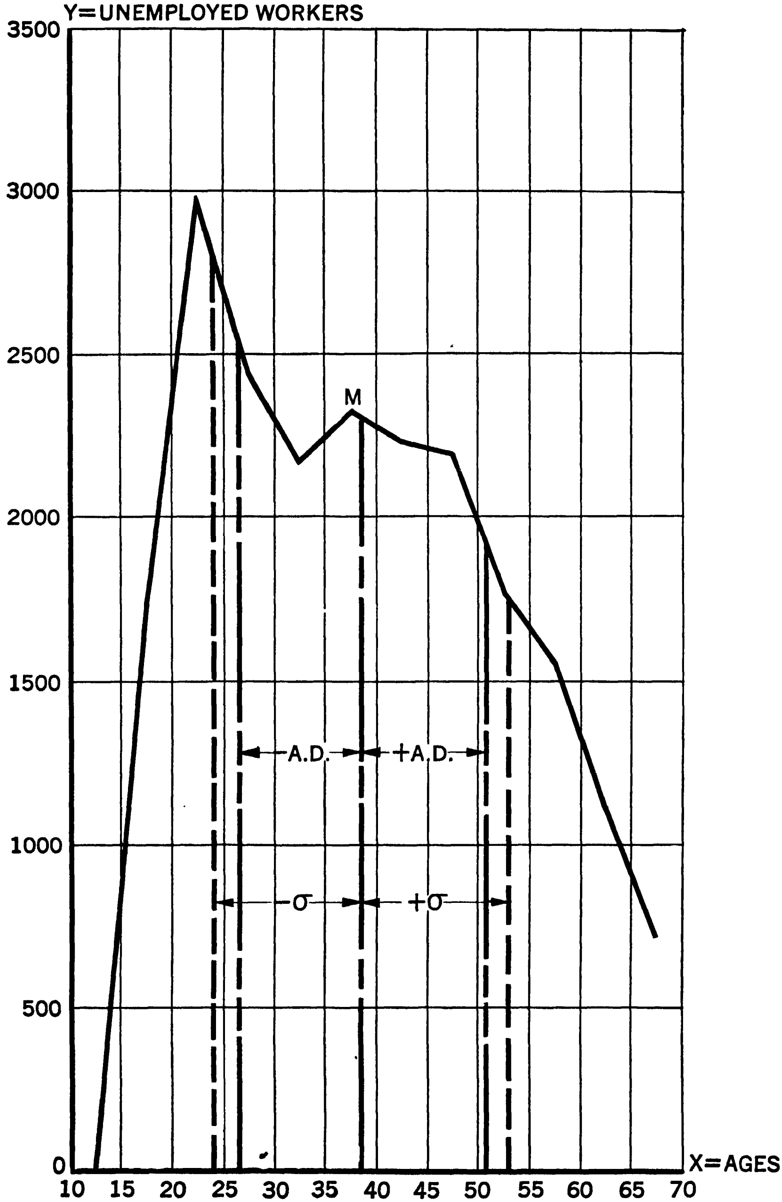


FIGURE LIII.—AREAS OF SURFACE ENCLOSED BY PLUS AND MINUS ONCE THE AVERAGE DEVIATION AND BY PLUS AND MINUS ONCE THE STANDARD DEVIATION FROM THE MEAN AGE OF BOSTON WORKERS

median plus the average deviation. Similarly, in a perfectly symmetrical distribution 68.26 per cent of all the items are included between the value equal to the mean minus the standard deviation and the value equal to the mean plus the standard deviation. The corresponding percentages in asymmetrical distributions will differ in varying amounts from these quantities for a normal distribution. In a normal distribution the values equal to plus and minus twice the standard deviation from the mean will include approximately 95.5 per cent, and the values equal to plus and minus three times the standard deviation from the mean will include approximately 99.7 per cent of all items. In asymmetrical distributions the percentages will vary, but it is a good rule to remember that the above percentages hold for ideal distributions.

7. COEFFICIENT OF RELATIVE VARIABILITY

The measures we have been discussing are measures of absolute variability. Sometimes, however, it is desirable to compare the variability of two statistical series expressed in different units of measurement. For example, we might want to express the comparative variability of wages expressed in weekly amounts and salaries expressed in monthly amounts. Obviously, the standard deviations of the two series would not be comparable. Some way must be found for expressing the relative variability of these two quantities. The required measure of relative variability will be the ratio of the measure of absolute variability to an average. In order to express the ratio as a percentage, it may be multiplied by 100.

There are several ways of computing the coefficient of relative variability, depending upon the measure of absolute variability and the type of average used. The formulas for computing the coefficient of relative variability are as follows:

$$V = \frac{\sigma}{M}$$

$$V = \frac{\text{A. D.}}{Md,\ M, \text{ or } Mo}$$

If the average deviation is used, the coefficient of relative variability may be computed with the use of the median, the mean, or the mode, but the same average should be used in this formula as was used in computing the average deviation. The use of these two formulas will be illustrated below, using the data for Boston unemployed workers:

$$V = \frac{\sigma}{M} = \frac{14.2}{38.7}$$
$$= .367, \text{ or } 36.7 \text{ per cent}$$

Using the average deviation from the median, instead of the standard deviation from the mean, the substitution is as follows:

$$Md \quad 37.8$$
$$= .320, \text{ or } 32.0 \text{ per cent}$$

There is no particular advantage in changing the ratio to a percentage except that we are more accustomed to thinking in terms of percentage. On the basis of the standard deviation, which is the most common way of computing the coefficient of relative variability, the coefficient of relative variability is 36.7 per cent. The ages of unemployed workers in some other city might be taken for purposes of comparison and the coefficient of variability computed to see whether there was less or more variability in the other city. A low coefficient of relative variability, like a low measure of absolute variability, indicates a high degree of homogeneity in the data for the trait measured.

8. MEASURES OF SKEWNESS

Up to this point the discussion of variability has been concerned with the individual items—the average variation of each item from some measure of central tendency. But sometimes it is desirable to have a measure of the variability of the whole mass of data. Previous measures of variability have not indicated the direction in which variability is most pronounced—that is, toward the lower or the higher values. The measure of this type of variability is called a measure of skewness. When data are plotted in frequency curves, they may be concentrated at one end or the other of the distribution—that is, the distribution may be skewed, as most empirical frequency distributions are. Hence, a measure of skewness shows the amount of skewness and the direction of the skew. Looking at Figure LIII, it is obvious that there is a concentration of ages at the lower end of the scale and that the tail of the curve is longer in the direction of the high age groups, which means that the age distribution is skewed in that direction.

Skewness is a function of both central tendency and variation from central tendency. Wherefore, it should be measured in terms

of these quantities. Two formulas are commonly used for computation of skewness:

$$Sk = \frac{M - Mo}{\sigma}$$

in which M is the arithmetic mean and Mo the mode. This is Karl Pearson's formula. The other formula is:

$$Q_1 + Q_3 - 2\ Md$$

Substituting in the first formula to find the skewness in the distribution of Boston unemployed workers, we have:

$$Sk = \frac{38.7 - 36.0}{14.2}$$
$$= +.19$$

This distribution is, thus, skewed slightly in the direction of the higher values. The mode was computed by the formula:

$$Mo = M - 3(M - Md)$$

The skew may vary from 0 to $\pm$ 1 but can never exceed 1.

9. EXERCISES

1. The following table gives the number of unemployed male workers in Chicago:

TABLE LII

UNEMPLOYED MALE WORKERS IN CHICAGO AT THE TIME OF THE CENSUS IN APRIL, 1930, ACCORDING TO AGE. CLASS A[1]

Age in Years	Number Unemployed
Total	122,685
10–14	19
15–19	9,399
20–24	18,283
25–29	15,686
30–34	13,870
35–39	15,014
40–44	13,996
45–49	12,602
50–54	9,439
55–59	6,790
60–64	4,784
65–69	2,803

[1] *Unemployment Bulletin, Illinois,* by the United States Bureau of the Census, 1930.

(a) Find the quartile deviation of the above age distribution.
(b) Find the average deviation of the above age distribution.
(c) Find the standard deviation of the above age distribution.
(d) Find the coefficient of relative variability for the above distribution.
(e) Find the coefficient of skewness for the above distribution.
(f) Compare your measures of dispersion with the measures of dispersion for the Boston unemployed men. Are there significant differences? If so, how do you account for them?

2. Consult the United States Census of 1930 concerning marital status in your own state. Compute by counties the per cent of the total population which is married. What is the standard deviation of these percentages? Do the same thing for another state in a different geographical section of the country. What differences do you find? How do you account for them? Can you see any way by which the differences in percentages married in different counties might affect such social problems as crime and dependency?
3. The following table gives the ratio of males per 100 females admitted to hospitals for the insane in 1927:

TABLE LIII

RATIO OF MALES PER 100 FEMALES ADMITTED TO HOSPITALS FOR THE INSANE BY STATES IN 1927[1]

State	Males per 100 Females
United States	140.4
Alabama	101.9
Arkansas	134.9
California	179.5
Connecticut	138.7
District of Columbia	295.3
Florida	142.7
Georgia	108.0
Illinois	167.7
Indiana	123.8
Iowa	144.9
Kansas	148.2
Kentucky	141.3
Maine	134.9
Maryland	127.9
Massachusetts	113.1
Michigan	183.1
Minnesota	151.0
Mississippi	149.6
Missouri	135.1

TABLE LIII—*Continued*

State	Males per 100 Females
Nebraska	164.0
New Hampshire	102.0
New Jersey	129.8
New York	128.2
North Carolina	117.9
Ohio	142.3
Oklahoma	139.6
Oregon	196.5
Pennsylvania	123.0
Rhode Island	138.0
South Carolina	104.2
Tennessee	123.7
Texas	111.8
Virginia	141.3
Washington	186.1
West Virginia	128.3
Wisconsin	185.3

Mental Patients in State Hospitals, United States Bureau of the Census, 1930. A few states are omitted, because rates are not given.

(a) Compute the 10th, 25th, 40th, 50th, 60th, 75th, and 90th percentiles for the ratios in the above distribution.

(b) Is there any way to account for the wide variation in the ratios? What about differences in administrative policies, differences in racial or national composition of the population, or the sex ratio in the states?

10. REFERENCES

Chaddock, Robert E., *Principles and Methods of Statistics*, Chap. IX.

Kelley, T. L., *Statistical Method*, Chap. IV.

Mills, Frederick C., *Statistical Methods*, Chap. V.

Thurstone, L. L., *The Fundamentals of Statistics*, Chaps. 13-16.

Walker, Helen M., *Studies in the History of Statistical Method*, Chaps. II (sec. 5) and IV.

CHAPTER X

Index Numbers

1. THE NATURE OF INDEX NUMBERS

AN INDEX number is a device for showing the average percentage change in prices, production, dependency, crime, etc., from one point of time to another or the variation from one geographical locality to another. An index number is, therefore, a kind of average, but it is so different from other averages that it is treated separately. Index numbers may be expressed as ratios or in terms of thousands, but generally they are expressed as percentages. It is difficult for the mind to grasp the relative size of crude quantities, but comparison becomes easy if the crude quantities are expressed as percentages of one of the quantities taken at a particular time or in a certain locality. Some period of time or geographical area is selected as the *base* to which quantities from all other periods are related in terms of percentage. The base year, month, or area is not selected carelessly; it serves best when it is about an average time or place. This base, then, becomes a sort of arbitrary "normal." As time passes it may be desirable to change the base period, because the original may cease to be representative or one nearer to the present time may be more satisfactory.

An illustration will make clearer the value of index numbers. The United States Bureau of Labor Statistics publishes an index number for the cost of living. This is concerned with what it costs families to live at one period as compared with a base period, and covers food, clothing, rent, fuel and light, house furnishing goods, and miscellaneous items in the family budget. The average cost of living in 1913 is taken as the base period and is denoted as 100.0. The average cost of living in each subsequent half-year is expressed as a percentage of the cost of living in 1913. According to the Bureau of Labor Statistics, the cost-of-living index in June, 1920, was 216.5. That is, in seven years' time there had been an

increase of 116.5 per cent in the cost of living. By the same standard the cost-of-living index in December, 1930, was 160.7.[1] The cost of living had declined markedly since 1920, but it was still 60.7 per cent higher than in 1913. If a family had retained its 1913 standard of living, its money income would have to be 60.7 per cent greater in 1930. This index is for cities and probably does not reflect exactly the cost of living in rural areas. Professor Paul H. Douglas computed an index of "real wages" from 1890 to 1926—"real wages" refers to the comparative purchasing power of wages at different periods. He found that in industry as a whole in the United States, using 1914 as a base of 100.0, the index for 1926 was 130.0.[2] The index of the cost of living computed by the Bureau of Labor Statistics stood at 174.8 in June, 1926. These two indexes are not quite comparable, because the cost-of-living index uses 1913 as a base and the real-wages index uses 1914. But even if 174.8 is a few points too high, it is clear that wage rates had not gone up as rapidly as the cost of living; consequently, there must have been a reduced standard of living among wage workers. If costs of living of rural people had been included in the cost-of-living index, it would be somewhat lower still, but making due allowance for this fact, up to 1926 the cost of living seems to have advanced more rapidly than real wages. This illustration shows the usefulness of an index number. It makes comparisons easy, because the relative size of the quantities in different years is expressed in terms of percentage and because the base period preceded the World War and represented a time of fairly normal economic conditions. Using both index numbers, we get a rough idea of the trend in the standard of living among wage workers, a fact of great importance to social workers and to students of the social sciences.

There is one index number which has probably more general use than any other, and that is an index of the general price level. Its aim is to measure the changing purchasing power of money, and it is employed in any kind of study dealing with money costs over a period of time. Several general price indexes have been computed. For purposes of illustration, the Index of the General Price Level published by the New York Federal Reserve Bank will be used. Indexes of either wholesale or retail prices do not

[1] *Monthly Labor Review*, Vol. 32, No. 2, p. 214.

[2] Douglas, Paul H., *Real Wages in the United States, 1890-1926*, p. 205. Boston: Houghton Mifflin Co., 1930.

accurately reflect the general price level. Because of this fact, Mr. Carl Snyder, of the New York Federal Reserve Bank, undertook to compute an index which would take into consideration all aspects of price. His index contained four major groups of prices: wholesale commodity prices, retail commodity prices, wages, and rents.[3] This index uses 1913 as the base year, or 100.0, and it includes annual indexes from 1875 to the present time; monthly indexes are also published. According to Snyder, the index of the general price level in 1920 was 193.0. That is, what a dollar would purchase in 1913 would take $1.93 in 1920. By 1930 the index had dropped to 168.0, that is, prices had fallen; or, to put it another way, the purchasing power of money had risen again. Any comparison of money costs from one year to another requires the use of a price index to reduce the volume to comparable dollars. For example, if the operation of a hospital cost $1,000,000 in 1913 and the same standards of service are maintained without effecting economies anywhere, the amount required in 1930 would be $1,680,000.

By this time it will have occurred to the student that the computation of an index of the cost of living or an index of real wages is complicated and laborious. The computation of some index numbers is much simpler, because fewer quantities are combined. Wherever many quantities have to be combined the process is long. Even in the food item of the cost-of-living index there enters the problem of combining costs of many kinds of foods. A means of assigning relative importance to these items of food has to be found. Then the relative importance of food, clothing, rent, etc., has to be determined before they can be combined to compute a general index of the cost of living. Methods of doing this will be described later in the chapter.

2. THE PRINCIPLE OF INDEX NUMBERS APPLIED TO SOCIAL DATA

Index numbers were invented as measures of changes in prices, but in recent years they have been applied to many other kinds of data. The Standard Trade and Securities Service publishes a compilation of several hundred index numbers. Some are general indexes, such as indexes of general prices, but many of them are specific indexes, such as indexes of prices of particular commodities or production in special industries. The application of the prin-

[3] Snyder, Carl, "The Measure of the General Price Level," *Review of Economic Statistics,* February, 1928, p. 10.

ciple of index numbers to sociological data is quite recent and not far developed, except in certain fields which lie on the border between strictly economic territory and the sociological field. These marginal fields are represented by indexes of the cost of living and of real wages. Furthermore, up to the present index numbers have dealt largely with time series. But there is no reason why they cannot be applied to many kinds of sociological data and to non-temporal series.

3. THE USE OF INDEX NUMBERS IN TIME SERIES

As stated above, the principle of index numbers was first applied to time series, particularly to price changes over a period of time. But to what kinds of sociological data can the principle be applied? The answer is that it can be applied to any kind of quantitative data which change in time. For a number of years Dr. Ralph G. Hurlin, of the Russell Sage Foundation, has been collecting data from family relief agencies, and he has worked out monthly indexes.[4] These show the changing case loads of reporting relief agencies month by month. A glance at the charts given by Dr. Hurlin is sufficient to see how the case load varies at different times of the year. For the monthly indexes January, 1926, is taken as the base period, or 100.0, and the case load of each succeeding month is expressed as a percentage of this period. The present writer has employed the principle of index numbers to measure the trend of the volume of public welfare work in Indiana.[5] Special indexes were computed for the number of persons aided per 100,000 population each year from 1900 to 1927 for each general type of public welfare work, including hospitals for the insane, penal institutions, poor asylums, child wards of the state, institutions for the feeble-minded, etc. A system of weights was devised, based upon the annual cost per person aided for each type of work, and then all the series were combined to form a general index of public welfare work in Indiana. The base year was 1913. In this general index corrections have been made for changing population and for the changing value of the dollar.[6] The general index shows a general rise in the volume of welfare work carried on by the State of Indiana, even when due allowance

[4] Hurlin, Ralph G., "Indexes of Family Case Work Loads," *Survey*, February 15, 1928.

[5] See "Indexes of Public Welfare Work in Indiana," *Social Forces*, December, 1929.

[6] See Table XXVIII, p. 215, for the general index.

has been made for population and the purchasing power of the dollar.

Whenever interest centers in rates of change or directions of change in a series of social data, the principle of index numbers is a possible method to determine these facts. Birth and death rates may be expressed in terms of index numbers with a fixed base period. The increase in the number of apartments in a city year by year is an indication of shift from the family dwelling to a collective type of housing; an index number showing the rate of change might be of considerable value to the construction industry, to investors, to school authorities, and to students interested in the birth rate. An index number of the work certificates issued to children before the legal working age would indicate to the issuing authorities the changing tendencies of children to leave school as soon as possible or to remain in school longer. The specific types of time series to which the principle of index numbers may be applied is limited only by the requirements of the problem in hand.

The use of index numbers in connection with data distributed in space is less familiar than their use in time series, but some index numbers of the former kind have been constructed with promising results. It is more common to use ratios or rates for geographical areas. For example, death rates are computed for census tracts, cities, counties, and states. These furnish a means of comparing death rates, or, for that matter, the incidence of any other social problem. If an average death rate is taken as a sort of norm, then we have substantially an index number, though it may not be expressed as a percentage of the average rate, the latter corresponding to the base period. Whether or not it is desirable to transpose rates for spatial data is largely a matter for the judgment of the investigator. Two illustrations of index numbers based upon spatial data will be given.

Dr. C. Luther Fry made use of index numbers to express church attendance in 32 counties, where he studied this subject. He took Salem County, New Jersey, as the base, or 100.0, and expressed the "attendance interest ratios" of the other 31 counties as percentages of this base county. His index numbers vary from 43.7 in Pend Oreille County, Washington, to 191.3 in Monroe County, Georgia. Alongside of his index numbers for "attendance interest ratios" he has placed index numbers for the "membership ratios" in the counties. He computes the degree of correlation between

attendance interest and membership ratios and finds it very high. Thus, the computation of index numbers here is done partly to indicate the variation in each series, but also to provide a basis for computing a coefficient of correlation.[7]

Professor C. Horace Hamilton has made another use of an index: to measure the relative roughness of topography. In the published report of his study of the relation of topography to social development in certain counties of Virginia he has not indicated whether or not he adopted a base county to represent 100.0. But it is apparent that his figures lend themselves to conversion to the conventional forms of index numbers. He says: "The social development of that area [Appalachian Highlands] is limited by its topography more than by any other one factor. In making social studies of such mountainous areas or in planning institutional development in them, it is desirable to have an accurate method of measuring the influence of topography. The problem resolves itself into the construction of an index of topography which can be used in making correlations with various social and economic conditions."[8] Professor Hamilton took a topographical map and drew on it vertical and horizontal lines three-eighths of an inch apart, this distance being equivalent to 2.5 miles. In each county the number of times the horizontal and vertical lines crossed a 500-foot contour interval or a stream was counted. The total count for the county was then divided by one-hundredth of the number of square miles in the county. This quotient is his index of topography. He found some high correlations between his indexes and other social factors in the counties, which partly demonstrates the usefulness of his index. By selecting a base county, his indexes could easily be transposed into conventional index numbers which could be put in an array to show the range of variations in topography for all counties in Virginia.

The two illustrations above suggest how the usefulness of index numbers will depend upon the problem in hand, but there is little doubt that this type of index numbers can become of much greater value in the future.

If the principle of index numbers is going to be used in the study of a problem, the collection of the requisite data comes in for early attention. The purpose of the index number will deter-

[7] Fry, C. Luther, *Diagnosing the Rural Church*, p. 111. New York: George H. Doran Co., 1924.

[8] Hamilton, C. Horace, "A Statistical Index of Topography," *Social Forces*, Vol. 9, No. 2, pp. 204, 205.

mine the criterion for collection of data. No formula for the computation of an index number will yield reliable results unless the data collected are suitable for the purpose. The worker must carefully define his purpose at the beginning of his work, and it should be stated as concretely as possible. For example, Professor Paul H. Douglas undertook to compute an index of relative living costs in non-agricultural areas. An urban index is desired; that limits the collection of data to cities. But all the items entering into the cost of living of a family had to be considered, and it was necessary to determine the relative importance of food, clothing, rent, etc., in order to weight the expenditures for quantities used. Appropriate weights had then to be selected. But he found that the relative importance of different items in the family budget changes over a period of time; therefore, it was necessary to change the weights after a certain year in the series was reached. This fact came to light in the process of collecting data for the index.[9] No mechanical rule can take the place of logic. The investigator must take care to understand the degree of homogeneity he is obtaining in his data and must observe the changing importance of the factors involved. This statement suggests that the accuracy of any index number depends upon the judgments the investigator made in the early stages of his work, and that it is highly relative. That is a fact. The validity of an index number is determined in large measure by the technical skill and painstaking care of the investigator.

Most index numbers are based upon samples of data in a statistical universe and not upon all of the existing data. If the index number is to represent approximately the actual situation, the sample data must be representative of the statistical universe under consideration. This raises the question of random sampling. A random sample of data in a given field is such a selection of data as to eliminate as completely as possible all influences except chance. For example, a random sample of the distribution of library borrowers in a city could be made by taking every fifth name in an alphabetical index of the borrowers. A random sample of relief agencies in New York might be made in the same way, but it happens that there are a few large relief agencies and a great many small ones. A random sample based upon alphabetical arrangement of the names of the agencies might not adequately represent the whole relief field because many small agencies and

[9] **Douglas, Paul H., *op. cit.*, Chap. IV.**

possibly one large one would be included in the sample. The method of proportional sampling might be preferable, if the whole relief field is to be represented fairly in an index of relief in New York. That is, agencies would be consciously selected and not left to chance; the judgment of the worker would determine the relative importance of the relief agencies in the whole field and would accordingly select the agencies to be used. In the computation of index numbers this is probably the better method to pursue; that is, examine the field carefully and then choose the data which give proportional representation to all types in the field.[10]

The question of primary and secondary data arises in the construction of an index number just as it does in other statistical problems. If the investigator collects the original data, he knows by experience a good deal about the homogeneity and appropriateness of his material. But some of his material may be secondary. What, then, is he to do? He must make some inquiry into the method of collection of the data and estimate their appropriateness for his own project. Rarely does an investigator construct an index number from nothing but primary data. His prices are taken from published tables, his weights to be used in measuring the cost of living are taken from some independent investigation, or his dependency data are taken from published reports. He must have some understanding of how these data were gathered and what standards of accuracy were observed. The construction of an index number would often be far too expensive if only primary data were used. Secondary data are satisfactory, but they must be used critically.

4. TYPES OF INDEX NUMBERS

No effort is made in this chapter to discuss a wide variety of formulas but merely to illustrate a few of those which may be used most readily by the student. For extensive discussions of the validity of different formulas the student is referred to Fisher's *The Making of Index Numbers*, and to Professor Willford I. King's more recent book, *Index Numbers Elucidated*. In this chapter the elementary methods of constructing index numbers will be described.

The simplest form of comparison of quantities is the crude figures. The quantities are added and allowed to stand without

[10] For further discussion of this point, see King, Willford I., *Index Numbers Elucidated*, pp. 64-66. New York: Longmans, Green and Co., 1930.

reduction to relatives and without the use of weights. This in reality is not an index number, because by definition an index number shows relative change in magnitude. For purposes of illustration and comparison the same data will be used in all the formulas. The data will be the average amount of relief per allowance case given by three family relief agencies of New York City during a period of four years, 1927 to 1930.

TABLE LIV

AMOUNT OF RELIEF PER ALLOWANCE CASE IN THREE NEW YORK FAMILY RELIEF AGENCIES[1]

Agency	Relief per Case, 1927	Relief per Case, 1928	Relief per Case, 1929	Relief per Case, 1930
No. 1	\$ 44.85	\$ 41.45	\$ 44.45	\$ 46.11
No. 2	49.29	49.54	51.00	52.95
No. 3	47.90	52.76	53.49	53.97
Total	\$142.04	\$143.75	\$148.94	

[1] From data compiled by the Department of Statistics of the Russell Sage Foundation. Indexes computed for these relief agencies might just as well have been computed in terms of case load; this would remove the changing price factor, and it would represent volume of work just as well.

Examination of the column totals reveals the fact that there has been an increase in the amount of relief per allowance case, but it is difficult to get a definite conception of the amount of change from year to year. The crude figures are too large, and they are not in any way related to each other. It is possible to make comparisons between the annual totals, but the percentage change can only be guessed. We need the totals expressed in some form that reveals the relative amount of relief per allowance case.

The simplest form of an index number consists of relatives based upon the sum of aggregate values unweighted. The formula for this index number may be expressed as follows:

I = index number for the given year
Σq_0 = sum of the quantities in the base year
Σq_1 = sum of the quantities in the given year

If there is only one quantity in each year, then the summation sign is omitted from the formula. For example, if Agency No. 1 were the only agency being considered, there would be no summation. But in Table LIV there are three quantities. The totals

of the columns, then, will be used in the formula, as follows, using 1927 as the base year, or q_0

$$I = \frac{143.75}{142.04}$$
$$= 101.2, \text{ index for } 1928$$

The indexes for the other years are 104.8 and 107.7, respectively. It is easy to grasp the significance of the changes in allowances, when reference is had to these index numbers. The increase in allowances over the base year was 1.2 per cent in 1928, 4.8 per cent in 1929, and 7.7 per cent in 1930. The sharpest rise occurred in 1929, but allowances are still going up. In view of the fact that the purchasing power of money was rising during this period, the increasing amounts of allowances appear to reflect a more liberal policy of relief giving. This might not be true in 1930, because the depression may have so depleted the slender resources of families that more relief had to be given for that reason. Whatever the explanation of the increasing amounts of allowances, the index numbers show that an increase is occurring, and that is their function.

Another method of computing an unweighted index for these data is that known as the average of relatives. The quantity for each agency in the base year is used as the base for computing relatives for that agency. Then the arithmetic mean of the relatives for each year is found. The variation in the formula may be expressed thus:

$$N$$

TABLE LV

AMOUNT OF RELIEF PER ALLOWANCE CASE IN THREE NEW YORK FAMILY RELIEF AGENCIES AND THE RELATIVES BASED UPON 1927

Agency	1927		1928		1929		1930	
	Relief	Relative	Relief	Relative	Relief	Relative	Relief	Relative
No. 1	$ 44.85	100.0	$ 41.45	92.4	$ 44.45	99.1	$ 46.11	102.8
No. 2	49.29	100.0	49.54	100.5	51.00	103.5	52.95	107.4
No. 3	47.90	100.0	52.76	110.1	53.49	111.7	53.97	112.7
Total	$142.04	300.0	$143.75	303.0	$148.94	314.3	$153.03	322.9
Average	47.35	100.0	47.92	101.0	49.65	104.8	51.01	107.6

Table LV shows how this type of index number is computed. The index numbers are substantially the same as when computed by the method of the sum of aggregates, though they are slightly higher by the method of average of relatives. If one or the other of the two preceding methods is to be used, the first is preferable because it requires less arithmetical work. In either case, a definite idea of the annual rising cost of allowances is made clear.

However, an examination of the table reveals the fact that the rising cost of allowances proceeds at different rates in the three agencies. This fact affects the index numbers as previously computed, but we are not sure that Agency No. 3 should affect the result as much as it does, or possibly it should affect it more. This result can be tested by devising a system of weights, based upon the number of allowance cases handled by each agency in the base year. Then, if the work of these three large relief agencies in New York can be assumed fairly to represent the policies of relief agencies in allowance cases, we shall have an index number reflecting the changing cost of allowance cases in the City of New York. This assumption may or may not be true; it would have to be tested by a study of some of the smaller relief agencies, but for purposes of illustration we shall make the assumption.

A weighted index number may be computed by the method of either the sum of aggregates or the average of relatives. Both methods will be illustrated for purposes of comparison, but first a system of weights must be determined. A convenient method of weighting the cost of allowances in this problem is to use the average monthly allowance case load in each agency, and then compute the percentage which each agency load constitutes of the sum of all the case loads. The following table shows this process:

TABLE LVI

AVERAGE MONTHLY ALLOWANCE CASE LOAD OF AGENCIES, AND THE WEIGHTS EXPRESSED AS PERCENTAGES OF THE TOTAL CASE LOADS

Agency	Monthly Case Load	Percentage of Total —the Weights
Total (1)	1,197 (2)	100.0 (3)
No. 1	304	25.4
No. 2	248	20.7
No. 3	645	53.9

TABLE LVII

COMPUTATION OF INDEX NUMBERS BY THE METHOD OF WEIGHTED AGGREGATES FROM THE ALLOWANCE CASE DATA

Agency	1927			1928			1929			1930		
No. 1	$44.85	25.4	$1139.90	$41.45	25.4	$1052.83	$44.45	25.4	$1129.03	$46.11	25.4	$1171.19
No. 2	49.29	20.7	1020.30	49.54	20.7	1025.48	51.00	20.7	1055.70	52.95	20.7	1096.07
No. 3	47.90	53.9	2581.81	52.76	53.9	2843.76	53.49	53.9	2883.11	53.97	53.9	2908.98
Total			$4742.01			$4922.07			$5067.84			$5176.24
Index			100.0			103.8			106.9			109.2

In this problem we shall use percentages as weights. The absolute numbers in column (2) could be used with about the same ease but, if these numbers were large, they would be cumbersome. In such cases percentage, or some other ratio indicating relative importance, is more convenient.

The following formula indicates the method of computing index numbers from weighted aggregates:

in which q_1 and q_0 have the same meaning as in the previous formula and W_1 and W_0 are the corresponding weights. Table LVII shows the method of computation.

The effect of weighting is to increase the size of the index numbers. If our weighting system is sound, it is evident that the unweighted index does not properly represent the changing amounts of allowances. When an index number is carried through a long period of years, the relative importance of its items often changes. When these changes become so large as materially to affect the results, the weighting system should be revised and applied from the point at which the changes became important. In this problem that could be determined each year by simply computing the percentage of the cases handled by each agency. Another way of determining the weights, when the index deals with data of past years, is to take the average annual percentage of cases carried by each agency for the entire period. Of course, when another year passes and the index number is computed for that year, either the old system of weights will have to be accepted as adequate or a new system computed and the index numbers revised for the entire period. It is perhaps easier to use different weights each year, based upon the annual allowance case loads of the agencies. Index numbers computed for other types of data require the same attention to weighting.

Another type of weighted index number is known as the average of relatives weighted. We shall illustrate the method of computing this kind of index number and compare the results with those obtained by the method of weight aggregates. The variation in the formula is as follows:

$$I = \frac{\sum\left(\frac{q_1 W_1}{q_0 W_0}\right)}{\Sigma W}$$

TABLE LVIII

COMPUTATION OF INDEX NUMBERS BY THE METHOD OF AVERAGE OF RELATIVES WEIGHTED FROM THE ALLOWANCE CASE DATA

Agency	1927			1928			1929			1930		
	Relative	W_0	Relative Times W_0	Relative	W_1	Relative Times W_1	Relative	W_2	Relative Times W_2	Relative	W_3	Relative Times W_3
No. 1..........	100.0	25.4	2540	92.4	25.4	2347	99.1	25.4	2517	102.8	25.4	2611
No. 2..........	100.0	20.7	2070	100.5	20.7	2080	103.5	20.7	2142	107.4	20.7	2223
No. 3..........	100.0	53.9	5390	110.1	53.9	5934	111.7	53.9	6021	112.7	53.9	6075
Total......			10,000			10,361			10,680			10,909
Index......			100.0			103.6			106.8			109.1

The relatives for individual agencies are taken from Table LV. Each relative is multiplied by its weight. The sum of the weighted relatives in each year is then divided by the sum of the weights, which is 100.0, and the resulting index numbers are almost identical with the results obtained by the method of weighted aggregates, as was to be expected. One method is as good as the other, but the method of weighted aggregates requires somewhat less arithmetical work.

At times it may be desirable, for special reasons, to shift the base year. If an index number extends over a number of years, conditions may so change that the original base year is unrepresentative of the period as a whole. In such cases the base year may be changed. If the index number has been constructed by the method of the average of relatives weighted, a good deal of recomputation is necessary to accomplish this. On the other hand, if the index number has been computed by the method of weighted aggregates, it is simple to shift the base year. All that is required is to select the new base year and then divide all the sums of aggregates by the sum for the new base year. For example, if it were desired to make 1929 the base year in the illustration given in Table LVII, we would simply divide 4742.01, 4922.07, and 5176.24 by 5067.84, and the new index numbers would be as follows: 1927, 93.5; 1928, 97.1; 1929, 100.0; 1930, 102.1. A change in the base year is equivalent to a change in the weights, because the relative size of the items in the new differs from their relative size in the old base year.[11] Hence, if it seems wise to shift the base year, a consideration of the weighting system is required, and new weights may have to be devised.

Index numbers may also be computed by the method of the geometric average of relatives or of aggregates. The nature of the geometric average is to show proportional differences. When it is used, the resulting index number is likely to be somewhat lower, except in the base year, than the index determined by the arithmetic average. The principal advantage of an index number in which the geometric average is used is that the base may easily be shifted. That will be illustrated by the problem which follows, and the formula may be written thus:

$$I = \frac{\sum \mathrm{Log}\left(\frac{q_1}{q_0}\right)}{N}$$

[11] See King, W. I., *op. cit.*, pp. 23-25, for a demonstration of this fact.

TABLE LIX

COMPUTATION OF INDEX NUMBERS BY THE METHOD OF THE GEOMETRIC AVERAGE OF RELATIVES FROM THE ALLOWANCE CASE DATA

Agency	1927		1928		1929		1930	
	Relatives	Logarithms	Relatives	Logarithms	Relatives	Logarithms	Relatives	Logarithms
No. 1	100.0	2.000000	92.4	1.965672	99.1	1.996074	102.8	2.011993
No. 2	100.0	2.000000	100.5	2.002166	103.5	2.014940	107.4	2.031004
No. 3	100.0	2.000000	110.1	2.041787	111.7	2.048053	112.7	2.051924
Total		6.000000		6.009625		6.059067		6.094921
Mean Log.		2.000000		2.003208		2.019689		2.031640
Index		100.0		100.7		104.6		107.6

The indexes as given in Table LIX are slightly different from those computed by other methods, but the differences are not great. However, these differences may be considerable. Suppose, now, that it is desired to shift the base to 1929. This is done by computing the relatives for the different years in terms of 1929 as the base, finding the logarithms for these relatives, and then taking the mean of the logarithms for each year. That is considerable work. The same results may be obtained, as Chaddock has pointed out,[12] by using the index 104.6 of 1929 as 100.0 per cent and dividing each of the other indexes in Table LIX by it. The resulting indexes on the new base are: 1927, 95.6; 1928, 96.3; 1929, 100.0; 1930, 102.9. If these indexes are plotted by the side of the indexes given in that table, it will be seen that the curves are parallel. That is, using the 1927 base, the ratio of the index for 1927 to the index for 1928 is .993, and, using the 1929 base, the ratio of the index for 1927 to the index for 1928 is .993. The geometric average shows proportional change, and the shifting of the base year does not affect the proportions of the index numbers when computed by the method of the geometric average of relatives.

The illustration just given is unweighted, but this average may be used equally well in the computation of a weighted index number. The logarithm of the relative is multiplied by the appropriate weight. The sum of the weighted logarithms for a given year, or other period, is divided by the sum of the weights. The quotient is the logarithm of the weighted index number desired.

5. THE "BEST" FORMULA

Much effort has been expended in trying to find an "ideal formula" for the construction of index numbers. Lately, however, less attention has been given to this question, and Professor King, one of the most recent writers on the subject, contends that there is no "best" formula.[13] The researches whose object was to discover an ideal formula may have an historical explanation which is to be found in the history of the uses to which index numbers have been thought applicable. In the beginning of the construction and use of index numbers the interest was almost exclusively in prices. An index number was synonymous with a measure of price

[12] Chaddock, *op. cit.*, pp. 185-187.
[13] King, *op. cit.*, pp. 219, 220.

variation. The "ideal formula" which has received the most attention is Irving Fisher's:

$$\sqrt{\frac{\Sigma p_1 q_0}{\Sigma p_0 q_0}} \times$$

The p's refer to prices for the base year and for another given year, and the q's refer to the quantities sold at the given price in the base year and in the other year considered. Obviously this ideal formula is in the old tradition of index numbers as measures of prices. The emphasis Fisher places upon index numbers as measures of prices shows his leaning toward the older conception of index numbers, though he distinctly states that index numbers may be used for other purposes. Nevertheless, he draws his illustrations for the numerous formulas from the field of prices. As long as prices furnished the data for index numbers, it was reasonable that a search should be made for a formula which would be "best" under all circumstances for handling this class of data. But when indexes of physical production, of dependency, of employment, of church attendance, etc., began to appear, the purpose of index numbers had so changed that it became apparent that the purpose of an index number, even when it deals with prices, should determine the formula.

The latter is the contention of Professor King. He points out that index numbers are means of answering specific questions about data. ". . . the nature of the question asked determines absolutely the mathematical procedure which must be used in arriving at the answer, in other words, no essential change in the method of solution is permissible except when the question to be answered changes."[14] In order to make King's position clear, as it applies to the data for allowance cases used above, we may restate two of his questions so that they apply to our data:

1. Considering that the work of each agency is of equal importance, what was the average ratio of allowances in 1928 to allowances in 1927?
2. How would the total amount of allowances in 1928 compare with the total allowances in 1927, if the same number of allowance cases had been handled in the two years?

The first question is answered by finding for each agency separately the ratio of the amount of allowances in 1928 to that in 1927 and

[14] *Op. cit.*, p. 26. See also pp. 51-56.

finding the average of the ratios for 1928. That is a simple arithmetic average of relatives unweighted. The answer to the second question is found by multiplying the mean allowance in each case for each agency in both years by the number of allowance cases for the agency in 1927. The products are added for each year, and then the ratio of the sum for 1928 to the sum for 1927 is found. This is the method of weighted aggregates. The two questions are different, and the answers are different.[15] Whenever an index number is required for a group of data, the first question to be asked is, not what formula to use, but what purpose the index number is to serve. When that is answered, the formula, or mathematical procedure, will be determined. As suggested above, various questions may be asked about the same data, and a different mathematical procedure is required to answer each. There is, then, no serious question of whether one formula is *per se* more accurate than another; the formula is correct if it answers the question asked.

The index numbers derived for the data on allowances by various methods differ more or less. These differences are shown in Table LX:

TABLE LX

Comparison of Indexes for Allowance Cases Computed by Different Methods

Year	Sum of Aggregates Unweighted (1)	Average of Relatives Unweighted (2)	Sum of Weighted Aggregates (3)	Average of Relatives Weighted (4)	Geometric Average of Relatives Unweighted (5)
1927.........	100.0	100.0	100.0	100.0	100.0
1928.........	101.2	101.0	103.8	103.6	100.7
1929.........	104.8	104.8	106.9	106.8	104.6
1930.........	107.7	107.6	109.2	109.1	107.6

The weighted indexes are somewhat higher than the unweighted indexes in each year above the base year, though the differences are not large. The differences between the unweighted index numbers is slight, and likewise the difference between the weighted index numbers. The unweighted index computed by the method of the geometric average is slightly smaller than either of the other unweighted indexes except for 1930, when it is the same as the index in column (2). True to one decimal place, the indexes

[15] *Ibid.*

for 1930 in columns (2) and (5) are the same, but, if carried to two decimal places, the one based upon the geometric average is slightly smaller. The geometric average minimizes extremes, and the effect is to give an index slightly smaller than other methods which utilize the arithmetic average or any other average except the harmonic mean. The use of the harmonic mean gives the lowest index of any of the averages. The differences between indexes computed by the above five methods will not always be as slight as they appear here. Consequently, the student should not conclude that it is a matter of indifference as to which one he uses. The one he selects for his use will depend upon what question he seeks to answer about his data.

A number of "tests" for the validity of index numbers have been proposed, but none has been entirely satisfactory, and King, as indicated above, maintains that as tests they are without merit—e.g., circular, factor-reversal, commodity-reversal, and time-reversal test. He would rest the validity of a formula upon the question of whether or not, when applied to the data, it answers the question asked. Truman L. Kelley has proposed the following tests for validity: the smallness of the probable error of the sample used, whether or not the results parallel habitual modes of thinking of the problem, proportionality of the index to the relatives, ease of entering or withdrawing items from the list of quantities used, ease of change of base period, and ease of change of unit of measurement in the list. On the basis of these tests Kelley finds that index numbers computed on the basis of the weighted geometric mean or the weighted median are the most reliable.[16] The so-called ideal formula proposed by Professor Fisher[17] requires complete data for its use; these are rarely obtainable. The principle laid down by King that the purpose of the index number determines the mathematical procedure seems to be as sound as any yet brought forward. Much more difficult to determine than the formula are the representativeness and adequacy of the sample of data. If these can be obtained and the purpose of the investigator is clearly stated, the formula is easily found.

6. EXERCISES

1. The following table gives the cost of maintenance of state institutions in Indiana from 1900 to 1930 inclusive:

[16] *Op. cit.*, pp. 341-347.

[17] See above, p. 318.

TABLE LXI

COST OF MAINTENANCE OF STATE INSTITUTIONS IN INDIANA, 1900–1930, IN ACTUAL DOLLARS[1]

Year	Cost	Year	Cost
1900	$1,290,790	1916	$2,794,867
1901	1,379,860	1917	3,016,533
1902	1,382,397	1918	3,228,806
1903	1,425,753	1919	3,306,288
1904	1,525,741	1920	3,748,893
1905	1,555,787	1921	4,026,403
1906	1,620,454	1922	4,049,277
1907	1,540,985	1923	4,173,881
1908	1,800,470	1924	4,154,984
1909	1,932,381	1925	4,600,119
1910	1,991,005	1926	4,544,566
1911	2,109,833	1927	4,765,332
1912	2,282,191	1928	5,060,151
1913	2,318,348	1929	5,145,641
1914	2,445,017	1930	5,392,771
1915	2,614,937		

[1] *Indiana Bulletin of Charities and Corrections*, July 1931, p. 353.

(a) Use an index of the general price level, such as that of the New York Federal Reserve Bank, and adjust the actual expenditures to comparable dollars. The Federal Reserve Index of the General Price Level can be found in the statistical reports of the Standard Trade and Securities Service.

(b) Make a graph showing the curves of actual dollars expended and the adjusted dollars expended.

(c) What was the percentage increase in expenditures between 1900 and 1930 in actual dollars? What was the percentage increase in expenditures in adjusted dollars?

(d) The population of Indiana at the time of the census from 1900 to 1930 was as follows: 1900, 2,516,462; 1910, 2,700,876; 1920, 2,930,390; 1930, 3,238,503. What was the per capita expenditure in each of these decennial years in actual dollars and in adjusted dollars? Has there been a marked increase in expenditures for the maintenance of state institutions during this period?

(e) What inferences might be drawn from the foregoing analysis of expenditures for maintenance of state institutions regarding the frequent complaints about the rising tax rate?

2. The following table gives the number of mental patients in

state hospitals of the United States at a number of different times between 1880 and 1928:

TABLE LXII

NUMBER OF MENTAL PATIENTS IN STATE HOSPITALS IN THE UNITED STATES IN SPECIFIED YEARS[1]

Year	Patients
1880	31,973
1890	67,754
1904	129,222
1910	159,096
1922	222,406
1923	229,664
1926	246,486
1927	256,858
1928	264,226

[1] *Mental Patients in State Hosptals*, United States Bureau of the Census, 1930, p. 6.

(a) Compute the number of mental patients in each year per 100,000 population of continental United States. The population will have to be estimated for intercensal years.
(b) Construct index numbers for the rates of mental patients per 100,000 population.
(c) Construct index numbers for the total patients each year without regard to changing population of the United States.
(d) Why do these two types of index numbers differ? What sort of question is answered by the one based upon rates? What sort of question does the other answer? Does the principle of weighting enter into either of these index numbers?

3. The next table gives the number of persons under care and the total cost of maintenance each year for the principal public welfare activities of the State of Indiana for a ten-year period, 1920 to 1929:
   (a) Compute an index number for public welfare work in Indiana by each of the five methods described in this chapter. Devise a weighting system that will give due importance to the different types of public welfare work.
   (b) In order to allow for changing population and changing purchasing power of the dollar it will be necessary to express the number of persons aided as the number per 100,000 population and to deflate the actual costs with an

TABLE LXIII

PERSONS UNDER CARE AND COST OF MAINTENANCE OF THE PRINCIPAL PUBLIC WELFARE AGENCIES AND INSTITUTIONS IN INDIANA, 1920–1929 [1]

Year	State Institutions		Poor Asylums		Dependent Children		Outdoor Relief	
	Persons	Cost	Persons	Cost	Persons	Cost	Persons	Cost
1920	11,505	$3,748,893	3,087	$1,085,349	4,462	$ 464,822	44,253	$ 417,230
1921	12,529	4,026,403	3,271	1,025,364	4,450	587,076	79,992	610,354
1922	12,937	4,029,277	3,365	1,021,941	4,487	612,628	94,850	741,174
1923	12,913	4,173,881	3,294	1,186,232	4,479	644,511	51,256	524,298
1924	13,949	4,154,984	3,301	1,113,469	5,456	733,897	71,725	618,902
1925	15,016	4,600,119	3,433	1,065,191	6,021	794,424	74,945	840,573
1926	15,769	4,544,566	3,535	1,197,831	6,367	776,611	93,302	972,082
1927	16,567	4,765,332	3,671	1,252,816	6,365	1,031,347	111,659	1,103,590
1928	17,211	5,060,151	3,969	1,353,081	6,984	917,317	126,711	1,274,674
1929	17,477	5,145,641	4,156	1,324,797	6,960	1,049,160	137,762	1,445,758

[1] *Op. cit.*, pp. 352, 353, 459. Persons and cost for outdoor relief estimated for 1926 and 1928.

index of the general price level. The population of Indiana in 1920 was 2,930,390; in 1930, 3,238,503.

(c) Show the five index numbers graphically on the same paper for purposes of comparison. Use the natural scale.

(d) What question is answered by the indexes based upon unweighted aggregates and upon unweighted relatives? By the unweighted index based upon the geometric mean? By the weighted indexes? From the point of view of public welfare, which question do you regard as the most important?

7. REFERENCES

Chaddock, R. E., *Principles and Methods of Statistics*, Chap. X.

Douglas, Paul H., *Real Wages in the United States, 1890-1926*, Chaps. IV, XIII, XXVIII, XIX.

Fisher, Irving, *The Making of Index Numbers*, Chaps. I-III.

Hurlin, Ralph G., "Indexes of Family Case Work Loads," *Survey*, February 15, 1928.

Kelley, Truman L., *Statistical Method*, Chap. XIII.

King, Willford I., *Index Numbers Elucidated*.

Mills, Frederick C., *Statistical Methods*, Chaps. VI, IX.

White, R. Clyde, "Indexes of Public Welfare in Indiana," *Social Forces*, December, 1929.

CHAPTER XI

Measurement of Relationships

I. THE CONCEPT OF CORRELATION

UP TO this point interest has centered in the description and analysis of a single series of data. A collection of social data presents a chaotic picture, until it is organized as an array or a frequency distribution. Something more is known about the data when an average is computed, and still more is known when the variation of individual items from the average is found. The method of index numbers makes possible a comparison of the magnitude of variables at different times or localities. Measures of central tendency and of dispersion have defined more precisely the frequency distribution, but they have given us no conception of the relationship between two or more series of social data. Sorokin has defined sociology in these words: "It seems to be the study, first, of the relationship and correlations between various classes of social phenomena (correlations between economic and religious; family and moral; juridical and economic; mobility and political phenomena and so on); second, that between social and non-social (geographical, biological, etc.) phenomena; third, the study of the general characteristics common to all classes of social phenomena."[1] If this concept of sociology is accepted, it is obvious that the social statistician is especially interested in the interrelationships of social phenomena. It is no less true of social work than of sociology; the central interest of the social worker is in the relations of different social factors to the condition or situation with which he deals. At this point, then, in the study of statistical methods it is appropriate to introduce ways of measuring relationships.

The study of relations is not peculiar to the social sciences. Relation is the paramount fact of all science. For example, the freezing

[1] Sorokin, Pitirim A., *Contemporary Sociological Theories*, pp. 760, 761. New York: Harpers, 1928.

point of water at 0° Centigrade is a measure of the relation between the condition of water and temperature at sea level. The symbol, H_2O, indicates the relation existing between definite quantities of hydrogen and oxygen under specified conditions. The so-called laws of physics, chemistry, and biology are statements of relationships. In view of this fact, it is less surprising that relationships in the social sciences should be regarded as paramount and that ways of measuring these relationships should occupy much of the attention of social scientists.

The traditional conception of cause is not used much in statistics. Cause-and-effect have had a history too closely connected with the older metaphysics to make them of use in the social sciences, unless the concept be redefined. The measurement of relations by statistical methods is the modern substitute for the metaphysical concept of cause-and-effect. Instead of speaking of one fact as a cause and another as an effect of the first, it is the habit to speak of one fact as the *independent variable* and of the second as the *dependent variable.* In some cases the dependent variable might just as well be treated as the independent variable. In other cases a certain amount of change in the independent variable is followed by a definite amount of change in the dependent variable. That approaches the traditional conception of cause-and-effect. Correlation is a method of measuring the degree of simultaneous variation existing between the averages and dispersions of a dependent variable and one or more independent variables. It may be a measure of cause-and-effect analogous to the traditional usage, but it is not necessarily so. If a change in one fact is so closely associated with change in another that the second may be predicted from the first, there is obvious interdependence which might be called a cause-and-effect relation, but, if so denominated, it should be clear that the relations are conceived in mechanistic terms as reactions to stimuli or forces. However, two facts may vary simultaneously and still not be related as independent and dependent variables. For example, the number of the population having tonsils removed may increase at the same time that the number of automobiles increases. The trends of the two series of facts might be correlated mathematically and an apparently significant coefficient of correlation found, but no one would assert that a cause-and-effect relation exists between the two sets of phenomena. Some understanding of the relation, if any, of two such kinds of phenomena is fundamental before inferences can be made regarding

cause-and-effect on the basis of statistical correlation. If there are good grounds for believing that two sets of phenomena vary interdependently, the technique of correlation may be employed to measure the degree of such interdependence. To state the matter another way: the discovery of a significant degree of correlation either confirms an hypothesis of interdependence or it suggests an hypothesis of interdependence requiring further consideration by other methods of analysis.

The technique of correlation is of particular importance in the study of social problems, because usually there is some social advantage to be achieved by obtaining control over the conditions among which social problems arise. If the aim is to reduce mortality in a certain area of a city, then the factors which contribute to a high mortality rate must be determined. Mortality here would be the dependent variable and the other factors the independent variables. In the first place, the independent variables with respect to mortality have to be identified. Then arises the question as to their relative importance as "causes" of mortality. This can be answered by the correlation technique in so far as covariation may be assumed to represent interdependence. Correlatedness there is between two series of social data, such as the interrelatedness of physical production and volume of employment or the age distribution of the population and the per cent of the population married. It should always be kept in mind that "correlation" in statistical discussion refers *only* to the degree of relations among *numerical variables*. If qualitative data are to be analyzed by the correlation technique, then they must be reduced to quasi-quantitative terms by the use of a rating scale. This caution has more than ordinary weight in social statistics, because so many social facts thought to be of great importance are qualitative. Correlation technique is none the less important in the study of social problems, but great care is necessary in its application to specific data.

2. THE MEANING OF FUNCTION[2]

It is customary to speak of the independent variable as X and the dependent variable as Y. Or sometimes the independent va-

[2] For much of the detailed procedure which follows in this chapter the author is indebted to Ezekiel's *Methods of Correlation Analysis*, John Wiley & Sons, New York, 1930, the most comprehensive volume yet published on this subject.

riable is designated X_1 and the dependent variable as X_2. If there are two or more independent variables the X's are given appropriate subscripts to indicate the variable to which reference is made.[3] The Y variable is also known as a *function* of the X variable. All that this means is that Y is dependent upon X—that a variation in X is followed by a corresponding variation in Y. In a loose way, the variation in X might be called the cause of the variation in Y. When one variable is said to be the function of a second variable, it simply means that a variation in the second accompanies a variation in the first—that is, a variation in X accompanies a variation in Y. This mathematical language is precise in its meaning, and expresses a complex situation in a few words.

Table LXIV and Figure LIV illustrate the concept of function by means of data taken from physics. If a body moves at a uniform velocity, the distance traveled at any given second is equal to the product of the velocity and the time in seconds:

TABLE LXIV

DISTANCE OF A BODY FROM THE STARTING POINT, IF IT MOVES AT THE RATE OF 5 FEET PER SECOND, AT SPECIFIED SECONDS

Time in Seconds	Distance in Feet from ; Point
1	5
2	10
3	15
4	20
5	25
6	30
7	35
8	40

The diagonal line connecting the dots gives a picture of the distance traveled by the moving body at any specified second or fraction of a second. It expresses graphically the relation between time and distance. Such a diagram is the simplest way of indicating the functional relation of an independent and a dependent variable. It is obvious that a change in the X variable is accompanied by a definite change in the Y variable. We know from physics that

$$d = vt$$

[3] In this text the methods of multiple correlation, partial correlation, and part correlation are not presented. They are properly considered in an advanced course in statistics.

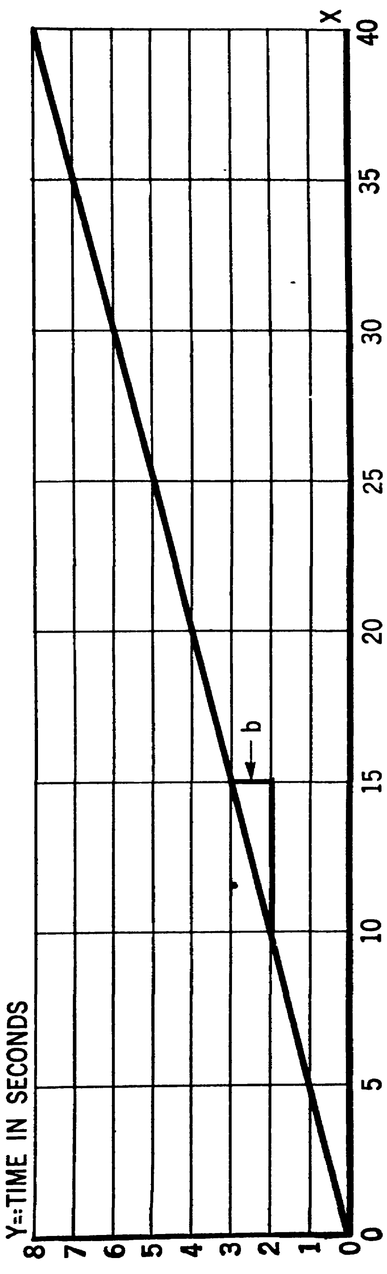


FIGURE LIV.—DISTANCE TRAVELED BY A BODY IN SPECIFIED TIME

in which d is distance, v velocity, and t time in seconds. Or, putting the formula in terms of X and Y,

or
$$Y = \frac{X}{v}$$

But suppose we think of the functional relation in terms of the diagonal straight line. This line of relationship may be expressed by an equation as well as by a graph. The general equation for a straight line is:

If the diagonal line had intersected OY above O, we should have had a small piece of OY below this intersection. That small piece of OY would be a in the above equation. The vertical distance b, indicated on the diagram, does not represent one second or a space on the diagram; it represents the ratio of seconds to feet traveled at any given second. The value of b must be computed. Since we know the value of a to be O, that will be easy to do by the method of simultaneous equations. It will be necessary to assume some value for X. It makes no difference what values are assumed for X; so for convenience we shall assume that the values of X at different times are 5 and 10. Since the velocity of the body is 5 feet per second, the corresponding values of Y will be 1 and 2. The equations may then be written as follows:

$$a + b(10) = 2$$
$$a + b(5) = 1$$

Or

$$a + 10b = 2$$
$$-a - 5b = -1$$
$$5b$$

$$5$$
$$b = .2$$

To solve the simultaneous equations, we assume the signs of the lower equation to be changed and then add algebraically, which cancels out the a's and leaves b as the only unknown. The value of b is found to be .2. Substituting the values of a and b in the equation for a straight line, we get:

$$Y = 0 + .2X$$

This is the equation of the diagonal line in the graph. It expresses the specific relation of X and Y. Every straight line relation between two series of data has a specific equation which may be found in the same manner as the above. Because the dots representing the intersections of ordinates and abscissas lay on a straight line, the equation makes exact estimates of unknown Y's possible. But many series of data, when plotted, are only approximately represented by a straight line. Then the specific equation which best "fits" the distribution is not so exact, and estimates made from it are only approximate. Such approximate equations are commonly found when we are dealing with social data.

As an example from social statistics, the relation between rates of misdemeanants and of felons in twenty census tract areas of Indianapolis will be used. Table LXV gives the data:

TABLE LXV
MISDEMEANANT AND FELON RATES BY CENSUS TRACTS, INDIANAPOLIS

Misdemeanant Rates X	Felon Rates Y
2.2	1.3
3.8	1.1
4.1	1.1
4.1	2.9
5.2	1.3
7.2	1.7
7.5	2.6
7.7	2.4
7.9	2.7
8.4	4.2
9.0	5.1
9.1	4.0
9.7	2.9
10.4	4.2
11.1	4.6
13.5	3.6
13.8	5.7
15.5	9.3
19.3	10.1
23.4	9.6

The relation between these two series of data is represented best by a straight line. Figure LV shows the data as a scattergram. There is considerable scatter among the dots, but they lie above and below a straight line which might be drawn through them.

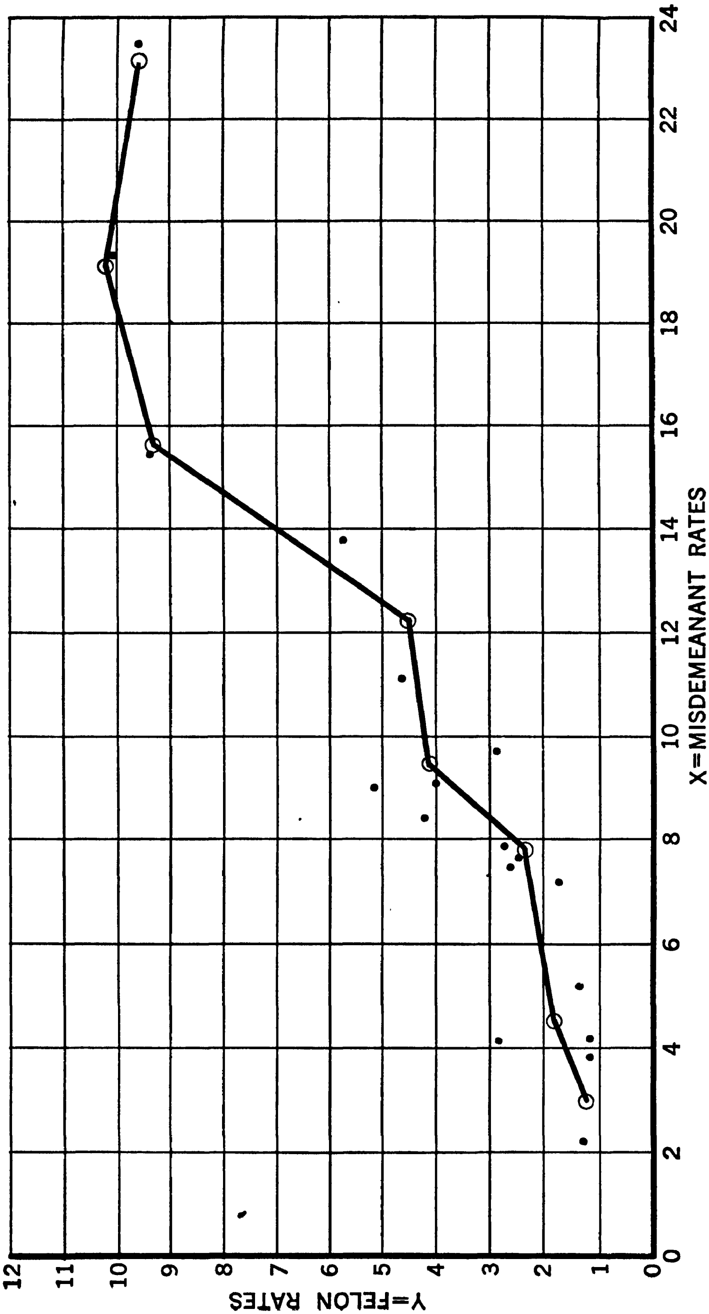


FIGURE LV.—MISDEMEANANT AND FELON RATES

The specific equation for the straight line must be found but, before computing the equation for the line of best fit, a line may be drawn free-hand which approximates the equation of the line. This is done by taking the mean rates for misdemeanants and felons in each class-interval and plotting them. The irregular line drawn through the dots is this line of means. It merely indicates a little more clearly the general relation between the two variables, felon and misdemeanant rates.

But another and more accurate method is required. The method commonly used is known as the method of least squares. The whole problem of fitting the line lies in determining the constants, *a* and *b*, in the equation for the straight line. The method seems to be a little complicated, but experience in using it soon dispels its apparently formidable character. When the student learns to use the method of least squares to fit a curve, he has learned a good deal of the procedure in the computation of coefficients and indexes of correlation. Table LXVI shows the computations necessary to determine the constants in the equation of the straight line.

TABLE LXVI

COMPUTATION OF VALUES (MISDEMEANANT AND FELON RATES) FOR DETERMINING THE LINE OF LEAST SQUARES

	Misdemeanant Rates X	Felon Rates Y	X^2	XY
	2.2	1.3	4.84	2.86
	3.8	1.1	14.44	4.18
	4.1	1.1	16.81	4.51
	4.1	2.9	16.81	11.89
	5.2	1.3	27.04	6.76
	7.2	1.7	51.84	12.24
	7.5	2.6	56.25	19.50
	7.7	2.4	59.29	18.48
	7.9	2.7	62.41	21.33
	8.4	4.2	70.56	35.28
	9.0	5.1	81.00	45.90
	9.1	4.0	82.81	36.40
	9.7	2.9	94.09	28.13
	10.4	4.2	108.16	43.68
	11.1	4.6	123.21	51.06
	13.5	3.6	182.25	48.60
	13.8	5.7	190.44	78.66
	15.5	9.3	240.25	144.15
	19.3	10.1	372.49	194.93
	23.4	9.6	547.56	224.64
Totals	192.9	80.4	2,402.55	1,033.18
Means	9.6	4.0	......	......

The following are the formulas for obtaining the values of a and b:

$$\Sigma XY - nM_xM_y$$

$$a = M_y - bM_x$$

In these equations M_x and M_y stand for the means of X and Y respectively, and n is the number of items. Substituting in these equations the values found from the table, we have:

$$\frac{\quad - \quad 768.00}{2,402.55 - 1,842.60} = .47$$

$$a = 4.0 - 4.51 = - .51$$

Putting these values in the place of the symbols, we have:

This is the equation of the straight line which fits the data on misdemeanants and felons and which describes the relation between

TABLE LXVII

VALUES OF Y ESTIMATED FROM VALUES OF X AND THE DIFFERENCE BETWEEN THE ACTUAL AND THE ESTIMATED VALUES

	Misdemeanant Rates X	Felon Rates Y	Values of Y Estimated from $Y = -.51 + .47X$	Residuals $(Y - Y')$	Residuals Squared
	2.2	1.3	.5	.8	.64
	3.8	1.1	1.3	− .2	.04
	4.1	1.1	1.4	− .3	.09
	4.1	2.9	1.4	1.5	2.25
	5.2	1.3	1.9	− .6	.36
	7.2	1.7	2.9	−1.2	1.44
	7.5	2.6	3.0	− .4	.16
	7.7	2.4	3.1	− .7	.49
	7.9	2.7	3.2	− .5	.25
	8.4	4.2	3.4	.8	.64
	9.0	5.1	3.7	1.4	1.96
	9.1	4.0	3.8	.2	.04
	9.7	2.9	4.1	−1.2	1.44
	10.4	4.2	4.4	− .2	.04
	11.1	4.6	4.7	− .1	.01
	13.5	3.6	5.8	−2.2	4.84
	13.8	5.7	6.0	− .3	.09
	15.5	9.3	6.8	2.5	6.25
	19.3	10.1	8.6	1.5	2.25
	23.4	9.6	10.5	− .9	.81
Totals	192.9	80.4	80.5		24.09

the two series of data. From it we can estimate the value of Y by assuming any value of X falling within the limits of the actual X's, that is, from 2.2 to 23.4. That the equation is not reliable for values of X outside the limits of the actual data is shown by the fact that, if we assume X to be 1, then the value of Y is —.04. It is nonsense to think of having less than 0 felons.[4]

Using the equation computed, we have estimated a value of Y for each value of X. The estimated values of Y are designated Y' (Table LXVII).

The total of the estimated values is close to the total of the actual values of Y. Thirteen of the estimated values are too large and 7 are too small. The straight line is, therefore, not a very close fit. However, it illustrates the method of fitting a straight line to data. If we had a larger number of cases, the fit might more evenly divide the minus and plus differences.

The residuals, or differences between the actual values of Y and the estimated values of Y', have been squared so that the *standard error of estimate* could be computed. The formula for the standard error of estimate is:

$$S^2_{y.x} = \frac{\Sigma z^2}{n - 1}$$

Substituting the values already obtained in this equation, we have:

$$S^2_{y.x} - \frac{24.09}{20}$$

Extracting the square root of both sides of the equation,

$$S_{y.x} = 1.1$$

The chances are approximately 2 to 1 that any estimate of the value of Y from the equation will not vary more than 1.1 above or below the actual Y value. The subscript of S indicates that Y is estimated from values of X; any value for X may be assumed, provided it is neither smaller nor larger than actual values of X given in the table.

The fitted straight line is shown in Figure LVI in relation to the actual distribution of data. The broken lines drawn parallel to the solid line represent the limits of the standard error of estimate above and below the estimated values represented by the solid line. The chances are 2 to 1 that any actual Y will fall between the broken lines.

[4] See Ezekiel, *op. cit.*, p. 60.

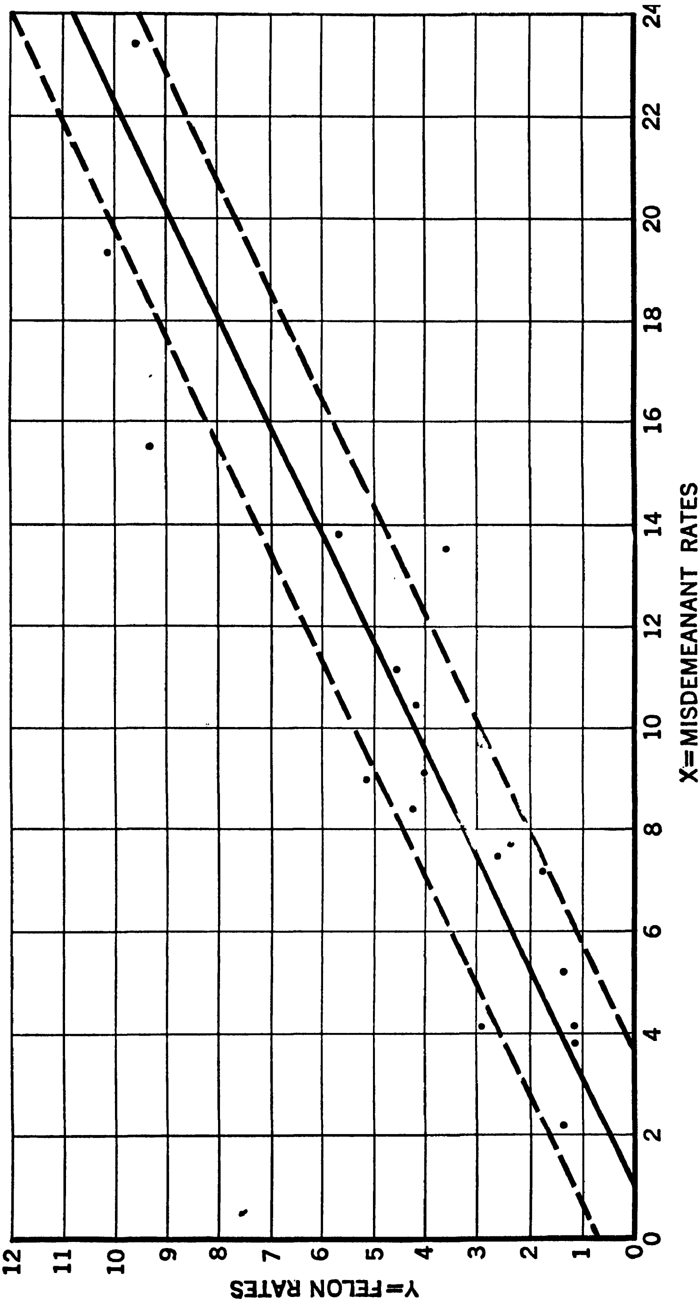
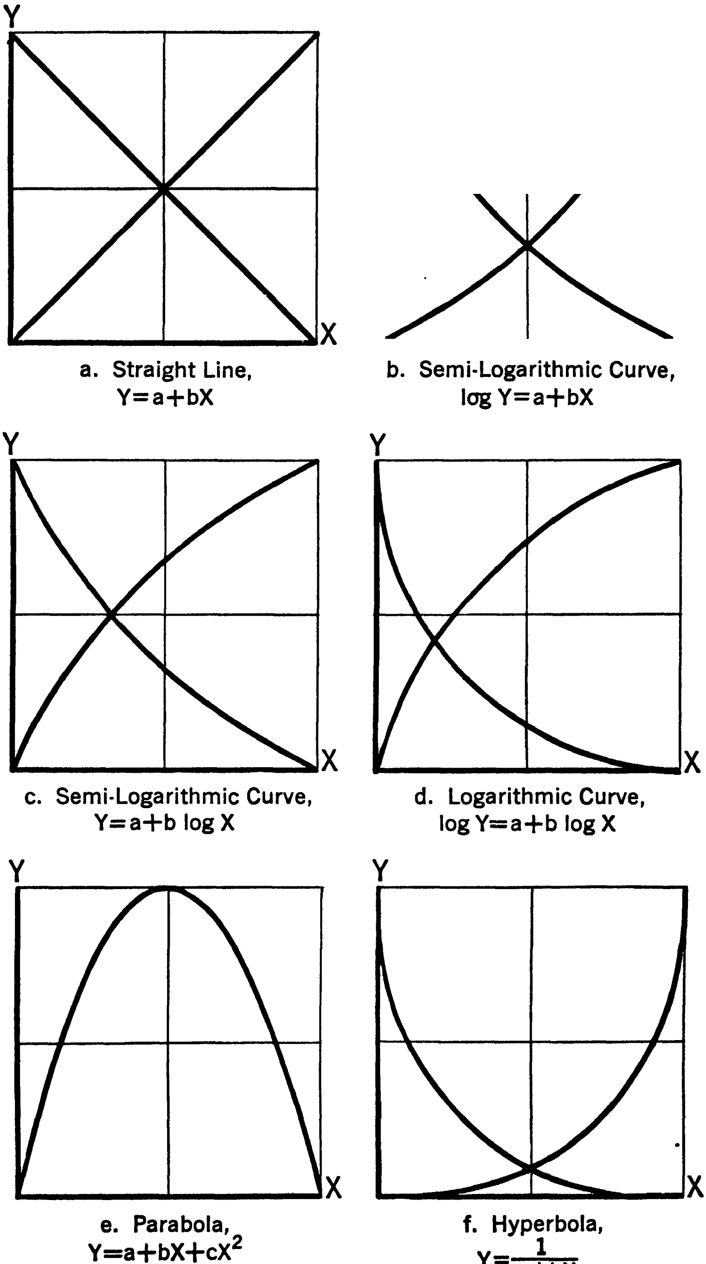


FIGURE LVI.—STRAIGHT LINE FITTED TO MISDEMEANANT AND FELON RATES



a. Straight Line, $Y=a+bX$

b. Semi-Logarithmic Curve, $\log Y=a+bX$

c. Semi-Logarithmic Curve, $Y=a+b \log X$

d. Logarithmic Curve, $\log Y=a+b \log X$

e. Parabola, $Y=a+bX+cX^2$

f. Hyperbola, $Y=\frac{1}{a+bX}$

FIGURE LVII.—TYPES OF STANDARD CURVES WITH THE FORMULA FOR EACH

The foregoing discussion related only to straight line relationships. But in the study of social statistics it will often be found that the relation between two variables is not linear, but curvilinear. The distribution may take the form of a parabola, a hyperbola, or a logarithmic curve. In such cases the fitted curve is computed by methods differing considerably from that used in fitting a straight line. Six common types of curves are shown in Figure LVII. It will be noticed that the formulas for curves involving the use of logarithms are identical with the formula for the straight line, except that the logarithm of X or Y or of both is used. Likewise the formula for the hyperbola in the lower right corner resembles the straight line formula, except that in the case of the hyperbola Y is equal to the reciprocal of $a + bX$. It is obvious that the fitting of a hyperbola is done in the same manner as fitting a straight line, and then the reciprocal for $a + bX$ is found, which is the value of Y. The fitting of a simple parabola involves the computation of a

TABLE LXVIII

PER CENT OF LAND USED FOR BUSINESS PURPOSES AND FELON RATE WITH COMPUTATIONS

	Per Cent of Land	Felon Rate	Logarithm of X		
			$\bar{X}$		$Y\bar{X}$
	7.1	2.6	.851	.7242	2.2126
	6.8	1.7	.833	.6939	1.4161
	4.6	2.7	.663	.4396	1.7901
	7.5	2.9	.875	.7656	2.5275
	16.3	5.7	1.212	1.4689	6.9084
	5.1	2.1	.708	.5013	1.4868
	5.3	1.6	.724	.5242	1.1584
	16.6	3.6	1.220	1.4884	4.3920
	13.3	5.1	1.124	1.2634	5.7324
	9.8	4.2	.991	.9821	4.1622
	9.0	2.5	.954	.9101	2.3850
	10.6	1.3	1.025	1.0506	1.3325
	15.9	1.1	1.201	1.4424	1.3211
	20.0	4.9	1.301	1.6926	6.3749
	23.3	4.0	1.367	1.8687	5.4480
	23.1	13.8	1.364	1.8605	18.8232
	22.0	4.4	1.342	1.8010	5.9048
	24.3	7.9	1.386	1.9210	10.9494
	33.9	2.8	1.530	2.3409	4.2840
	34.0	3.3	1.531	2.3440	5.0523
Totals	308.5	78.2	22.202	26.0834	93.6817
Means	15.4	3.9	1.110		

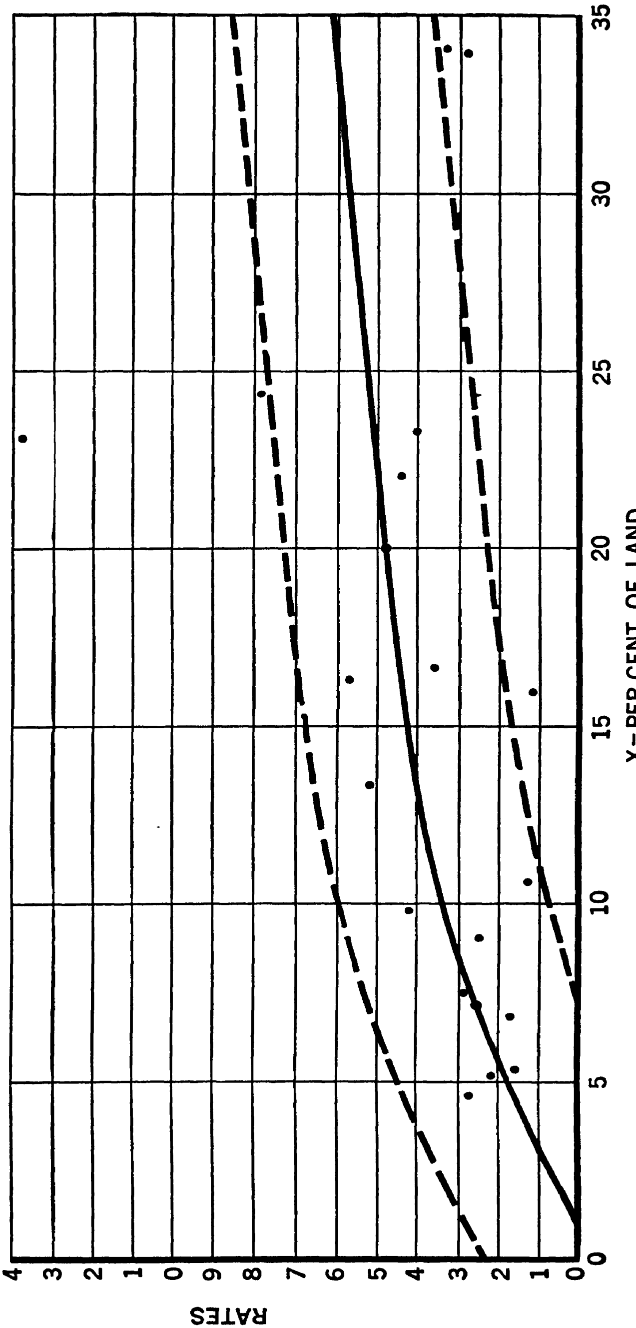


Figure LVIII.—Felony Data with Fitted Curve and Limits of Error of Estimate

third constant, c, and of X^2. In view of the fact that each of the logarithmic formulas is used in a similar manner, only one will be illustrated. A simple parabola will be fitted to the same data to illustrate the use of this formula.

The data used for the illustration are the per cent of land used for business purposes and the felon rate in certain census tract areas in Indianapolis. Table LXVIII gives the data and the necessary computations for fitting a logarithmic curve of the form $Y = a + b \log X$.

The first step is to find the values of the constants, a and b, and the formula is similar to that used in finding the constants for the straight line equation:

$$b = \frac{\Sigma Y\ \overline{X} - nM_yM_x}{\Sigma \overline{X}^2 - n(M_x)^2}$$

$$a = M_y - b\ M_x$$

The bar over X or Y indicates that the logarithm is used instead of the actual figures. Substituting in the above equations, we have:

$$b = \frac{93.6817 - 86.5800}{26.0834 - 24.6420} = 4.93$$

$$a = 3.9 - 4.93(1.110) = -1.57$$

The equation of the particular curve which fits the data is, therefore:

$$Y = -1.57 + 4.93\overline{X}$$

The equation is expressed in terms of Y. So, in using this equation for purposes of estimation it is not necessary to convert the values of Y to the anti-logarithms, or natural numbers; the values obtained will be actual values of Y.

Table LXIX gives the estimated values for Y and the residuals, that is, the differences between actual values and estimated values of Y.

Figure LVIII shows the distribution of the data, the fitted curve, and the limits of the standard error of estimate. The formula for the standard error of estimate for curvilinear distributions is similar to that for linear distributions. It is:

n —

TABLE LXIX

ACTUAL VALUES OF Y, ESTIMATED VALUES OF Y, AND THE RESIDUALS

Y	Y'	$Y - Y'$, or z	z^2
2.6	2.6	0	0
1.7	2.5	− .8	.64
2.7	1.7	1.0	1.00
2.9	2.7	.2	.04
5.7	4.4	1.3	1.69
2.1	1.9	.2	.04
1.6	2.0	− .4	.16
3.6	4.4	− .6	.36
5.1	4.0	1.1	1.21
4.2	3.5	.7	.49
2.5	3.1	− .6	.36
1.3	3.5	−2.2	4.84
1.1	4.4	−3.3	10.89
4.9	4.8	.1	.01
4.0	5.1	−1.1	1.21
13.8	5.1	8.7	75.69
4.4	5.0	− .6	.36
7.9	5.3	2.6	6.76
2.8	6.0	−3.2	10.24
3.3	6.0	−2.7	7.29
78.2	78.0		123.28

The standard error of estimate is 2.5. That is large. If the one extreme variation is left out, the standard error of estimate for 19 of the items is 1.5. Although the curve is not a very good fit, the logarithmic curve of this type would probably fit the data more closely if a larger number of items were included. The method of computation is the same, regardless of the closeness of fit.

While these data on crime are not distributed in the form of a simple parabola, we shall use them for purposes of illustrating the method of fitting a parabola.[5] As in the previous problem, the chief task is to compute the constants a, b, and c in the formula

$$Y = a + bX + cX^2$$

The a and b values will differ from their values in the logarithmic formula, unless c is zero. To determine the constants the following equations must be solved:

$$a = M_y - bM_x - cM_u$$

[5] See Ezekiel, *op. cit.*, pp. 72-78. Also F. C. Mills, *op. cit.*, pp. 284-290, 300-306.

To find the values necessary for the solution of these equations a table similar to the one used for the values for the logarithmic formula can be used. For convenience U may be used for X^2 in certain combinations which will be shown in the table. The method of determining the values of the above equations is as follows:

$$M_x =$$

$$M_y = \frac{1}{n}$$

$$= \Sigma X^2 \quad - n(M_x)^2$$

$$- nM_xM_y$$

$$- nM_uM_y$$

TABLE LXX

COMPUTATION OF VALUES FOR FITTING A SIMPLE PARABOLA—CRIME DATA

Per Cent of Land X	Felony Rate Y	X^2 or U	XU	U^2	XY	UY
7.1	2.6	50.41	357.91	2541.17	18.46	131.07
6.8	1.7	46.24	314.43	2138.14	11.56	78.61
4.6	2.7	21.16	97.34	447.75	12.42	57.13
7.5	2.9	56.25	421.86	3164.06	21.75	163.13
16.3	5.7	265.69	4330.75	70596.49	92.91	1513.43
5.1	2.1	26.01	132.66	676.52	10.71	54.62
5.3	1.6	28.09	148.88	789.05	8.48	44.94
16.6	3.6	275.56	4574.30	75955.36	59.76	992.02
13.3	5.1	176.89	2352.64	34931.61	67.83	902.14
9.8	4.2	96.04	941.19	9223.68	41.16	403.37
9.0	2.5	81.00	729.00	6561.00	22.50	202.50
10.6	1.3	112.36	1191.02	12633.76	13.78	146.07
15.9	1.1	252.81	4019.70	63907.84	16.49	278.09
20.0	4.9	400.00	8000.00	160000.00	98.00	1960.00
23.3	4.0	542.89	12649.34	294740.41	93.20	2171.56
23.1	13.8	533.61	12326.39	284728.96	318.78	7363.82
22.0	4.4	484.00	10648.00	234256.00	96.80	2129.60
24.3	7.9	590.49	14348.91	348690.25	191.97	4664.87
33.9	2.8	1149.21	38958.22	1320201.00	94.92	3217.79
34.0	3.3	1156.00	39304.00	1336336.00	112.20	3814.80
308.5	78.2	6344.71	155799.54	4262519.05	1403.68	30289.56

$M_x = 15.5.$ $\qquad M_y = 3.9.$ $\qquad M_u = 317.2.$

Substituting the required values in the preliminary equations shown above, we have:

$$\begin{aligned}\Sigma x^2 &= 6344.71 - 4805.00 = 1539.71\\ &= 155799.54 - 98332.00 = 57467.54\\ &= 4262519.05 - 2012316.80 = 2250202.25\\ &= 1403.68 - 1209.00 = 194.68\\ &= 30289.56 - 24741.60 = 5547.96\end{aligned}$$

Using these derived values in the equations to be solved simultaneously, we have

$$\begin{aligned}1b + 57467.54c &= 194.68 \quad \text{(I)}\\ 57467.54b + 2250202.25c &= 5547.96 \quad \text{(II)}\end{aligned}$$

These equations are most easily solved by the Doolittle method. Putting down equation (I), dividing it through by the coefficient of b with the sign changed, and placing the derived equation (I′) under it, we have

$$\begin{aligned}1539.71b + 57467.54c &= 194.68 \quad \text{(I)}\\ -b - 37.32361c &= -0.12643 \quad \text{(I}'\text{)}\end{aligned}$$

Equation (II) is then put down, equation (I′) is multiplied by the coefficient of c in equation (I), the result of which is placed under equation (I):

$$\begin{aligned}57467.54b + 2250202.25c &= 5547.96 \quad \text{(II)}\\ \underline{-57467.54b - 2144896.05c} &= \underline{-7265.62} \quad \underline{\text{(I}'\text{ times }57467.54)}\\ \text{Adding,}\qquad 105306.20c &= -1717.66\\ c &= -\ 0.01631\end{aligned}$$

Substituting this value of c in equation (I), we have

$$\begin{aligned}1539.71b + -937.30 &= 194.68\\ 1539.71b &= 194.68 + 937.30\\ b &= .74\end{aligned}$$

The third constant may now be computed:

$$\begin{aligned}a &= M_y - bM_x - cM_u\\ &= 3.9 - (.74)(15.5) - (-0.01631)(317.2)\\ &= 3.9 - 11.32 + 5.17\\ &= -\ 2.25\end{aligned}$$

The equation for the parabola is, therefore:

$$Y = -\ 2.25 + .74X - 0.01631X^2$$

With this equation the values of Y may be estimated for values

of X lying between the lowest and the highest actual values given in Table LXX. But the parabola does not fit these data, and the detailed estimates are not presented here. The logarithmic curve is a better fit.

3. MEASUREMENT OF THE DEGREE OF RELATIONSHIP

The methods of measuring relationship have so far shown whether or not a relation existed between the two series of data and have shown how closely the values of the dependent variable may be estimated from values of the independent variable. The first indication of relationship was determined by finding an equation which appeared to conform to the distribution and by plotting a curve with the estimated values. The second result was obtained by computing the standard error of estimate. Neither of these methods gives a clear idea of the importance of the interrelationship. Another method is necessary for this purpose.

This is the method of correlation. The degree of correlation is expressed as a coefficient. The degree of correlation in linear relations, that is, relations which may best be represented by a straight line, may vary from -1.0 to $+1.0$. If the coefficient comes out with a minus sign, it means that, when a change occurs in the independent variable, a corresponding change in the opposite direction occurs in the dependent variable. If the sign of the coefficient is plus, then a change in the independent variable is accompanied by a corresponding change in the same direction in the dependent variable. A coefficient of plus or minus one would be perfect correlation. In curvilinear relations the coefficient may vary from 0 to 1.0, the latter being perfect correlation. The measure of relationship in curvilinear relations is called an *index of correlation* and is designated by ρ to distinguish it from the measure of relationship in linear relations which is called a *coefficient of correlation* and is designated by r.

Perfect correlation is almost never found between two variables. A certain amount of variance in X is accompanied by a certain amount of variance in Y, and the measure of this interdependence is something less than unity. Social factors are influenced by a variety of things, and the aim of the statistician is to measure the amount of influence which an independent variable has upon a dependent variable or the amount of influence several independent variables have in combination upon a dependent variable. If a reliable measure of such relations can be obtained, the first step

has been taken toward control in the case of the variables considered. It cannot be said that a coefficient of correlation, if multiplied by 100, indicates the percentage of variance in Y due to variance in X. A slightly different measure is needed for this purpose. "Where both X and Y are assumed to be built up of simple elements of equal variability all of which are present in Y but some of which are lacking in X," says Ezekiel, "it can be proved mathematically that r^2 measures that proportion of all the elements in Y which are also present in X. For that reason in cases where the dependent variable is known to be causally related to the independent variable, r^2 may be called the *coefficient of determination.* It may be said to measure the percent to which variance in Y is determined by X, since it measures that proportion of all the elements of variance in Y which are also present in X." Likewise, "Where curvilinear relations have been used in determining the relationship, the term '*index of determination*' will be used to denote the value of ρ^2, thus retaining the same relation to the index of correlation that the coefficient of determination bears to r, the coefficient of correlation."[6] Hence we have two measures each for the degree of correlation in linear relations and in curvilinear relations, but they mean slightly different things. We shall illustrate, first, the method of computing the coefficient of correlation and the coefficient of determination for ungrouped data.

The formula for the coefficient of correlation used here is:[7]

= coefficient of correlation
= sum of the products of the two corresponding variables
nM_xM_y = product of the means times the number of items
ΣX^2 = sum of the squares of the X-variable
$nM_x{}^2$ = the square of the mean of the X-variable times the number of items
= sum of the squares of the Y-variable
= the square of the mean of the Y-variable times the number of items

The method of computing these values is illustrated by Table LXXI.

[6] *Op. cit.*, p. 120. [Italics mine. R. C. W.]
[7] This formula is taken from Ezekiel, *op. cit.*, p. 127.

TABLE LXXI

COMPUTATION OF VALUES FOR DETERMINING THE COEFFICIENT OF CORRELATION

Per Cent of Land X	Felon Rate Y	X^2	Y^2	XY
2.2	1.3	4.84	1.69	2.86
3.8	1.1	14.44	1.21	4.18
4.1	1.1	16.81	1.21	4.51
4.1	2.9	16.81	8.41	11.89
5.2	1.3	27.04	1.69	6.76
7.2	1.7	51.84	2.89	12.24
7.5	2.6	56.25	6.76	19.50
7.7	2.4	59.29	5.76	18.48
7.9	2.7	62.41	7.29	21.33
8.4	4.2	70.56	17.64	35.28
9.0	5.1	81.00	26.01	45.90
9.1	4.0	82.81	16.00	36.40
9.7	2.9	94.09	8.41	28.13
10.4	4.2	108.16	17.64	43.68
11.1	4.6	123.21	21.16	51.06
13.5	3.6	182.25	12.96	48.60
13.8	5.7	190.44	32.49	78.66
15.5	9.3	240.25	86.49	144.15
19.3	10.1	372.49	102.01	194.93
23.4	9.6	547.56	92.16	224.64
Total 192.9	80.4	2402.55	469.88	1033.18
Mean 9.6	4.0			

Substituting in the equation, we have:

$$r = \frac{1033.18 - 20(9.6)\,(4.0)}{5 - 20(92.16)]\ [469.88 - 20(16.0)]}$$

$$\frac{1033.18 - 768.00}{\sqrt{(559.35)\,(149.88)}}$$

$$= \frac{265.18}{289.54}$$

$$= .916$$

This is the unadjusted coefficient of correlation between the residences of misdemeanants and the residences of felons.[8] The stand-

[8] Another way to compute the coefficient of correlation is known as the product-moment method proposed by Karl Pearson. The formula for this method is

in which x and y are the deviations from the respective means. The same kind of table is used for computing the values as above, except that two additional columns are necessary for the deviations from the means. The product-moment method is somewhat longer than the method used above, and the results obtained are almost identical. Consequently, the method used for illustration is preferable as a labor-saving device.

ard deviation of either X or Y may be obtained from the values computed in the following manner: subtract the square of the mean multiplied by the number of items from the sum of the squares of X or Y, divide by the number of items, and extract the square root.

It was noted above that the coefficient .916 is unadjusted, that is, the number of items, or observations, has not been allowed for. That is particularly necessary when the number of items is small, as in the illustration. The following formula is used for adjusting the coefficient for the number of items:

The expression $(n - 2)$ stands for the number of items less the number of constants (a, b, c, etc.) in the equation describing the relation. Since the relation between misdemeanants and felons seems to be linear, there are two constants, because there are two constants in the equation of a straight line. Substituting the values in the equation, we have:

$$\bar{r}_{yx} = .911$$

Adjustment for the number of observations slightly reduces the size of the coefficient. Since the coefficient of determination is the square of the adjusted coefficient of correlation, we have

$$\bar{r}_{yx}^{2} = .830$$

In linear correlation the correlation may be either positive or negative. If the two series of data change simultaneously in the same direction, the correlation is positive. Examination of the table above shows that, when misdemeanant rates are high, felon rates are also likely to be high. To indicate the direction of variation, a plus sign may be placed in front of the coefficient: +.911. If one series of the data had shown a decrease when the other showed an increase, the correlation would have been negative, and a minus sign would have been placed before the coefficient. The required sign can always be determined by inspection of the table of variables or of a scattergram. On a scattergram if the dots are distributed in a rising direction from left to right, the correlation is indicated as positive. If the dots are distributed in a falling direc-

tion from left to right, the correlation is negative.[9] When the computation of the correlation is carried through, the coefficient comes out with the appropriate algebraic sign affixed; from the scattergram the sign can be guessed in advance of computation.

The regression equation and the standard error of estimate have not yet been computed. The regression equation requires the computation of the constants for the equation of a straight line,

$$Y = a + bX.$$

We shall compute the value of b, when Y is the dependent variable.

$$' - nM_xM_y$$

$$= \frac{1033.18 - 768.00}{2402.55 - 1843.20}$$

$$= \frac{265.18}{559.35} = .476$$

$$a = M_y - bM_x$$

$$= 4.0 - 4.570 = -.570$$

The regression equation of Y on X is, then,

$$Y = -.570 + .476X$$

Logically and practically, it is the estimation of Y from X that is desired. It is possible, however, to treat Y as the independent variable and X as the dependent variable in the regression equation and compute the change in X for each unit of change in Y from the following variation in the preceding formula:

Since the arithmetic involved in this formula is the same as in the preceding formula, the computation is not carried out.

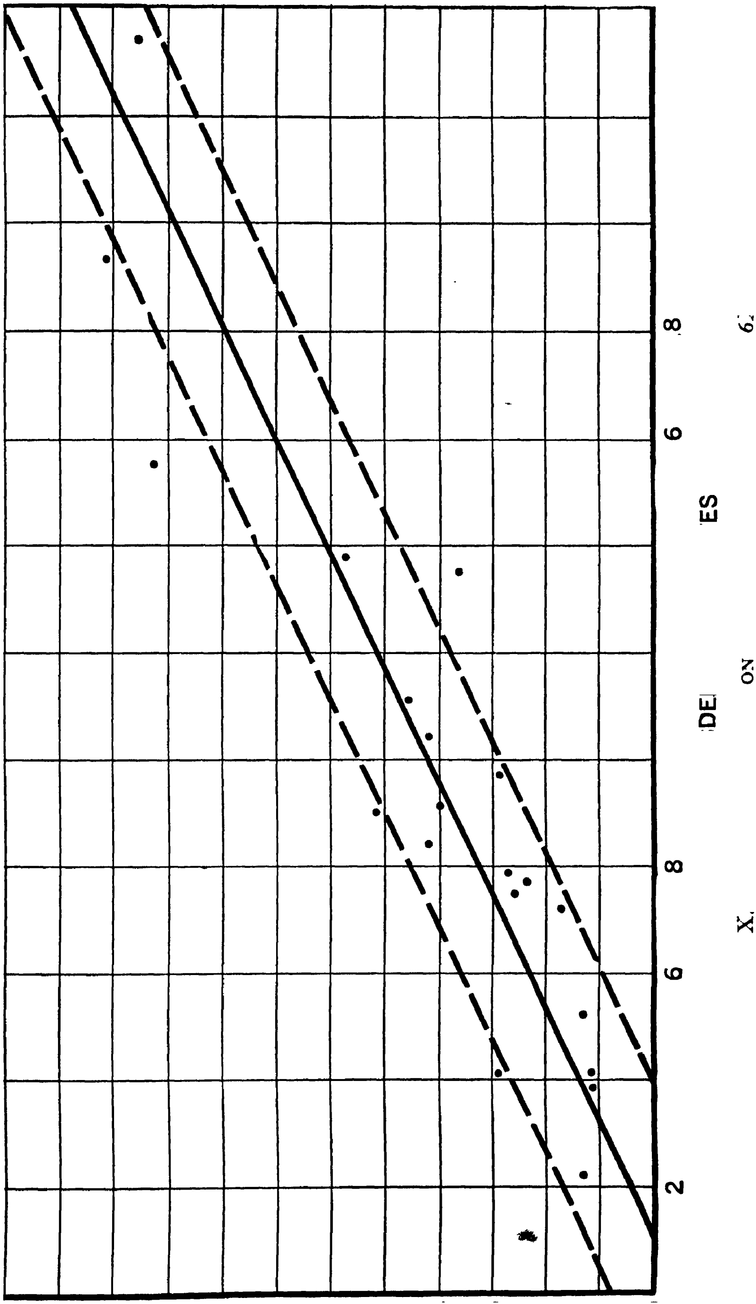
The formula for the standard error of estimate for the adjusted coefficient of correlation is as follows:

$$\bar{S}_{yx} = \qquad n - 1$$

Or, $\bar{S}_{yx} = \quad \sqrt{.60}$

$$\doteq 1.17$$

[9] **This is true, of course, only if the X-scale runs from left to right and the Y-scale from bottom to top. Reversal of the Y-scale would reverse the direction of the regression line.**



We may now make a scattergram, draw the regression line from estimates of Y, and insert the lines parallel to the regression line to show the limits of the range of the standard error of estimate. Figure LIX shows the regression line and the standard error of estimate determined from the coefficient of correlation.

The chances are 2 to 1 that any estimate of the value of Y from a value of X will fall within the limits indicated by the parallel broken lines. It is worthy of note that the standard error of estimate for the fitted straight line is 1.1, whereas the standard error of estimate determined from the coefficient of correlation is 1.16, or considerably larger than the first. The latter is more dependable, because it includes a consideration of the relative importance of the variations of the two variables.[10]

Up to this point the discussion of correlation has dealt with the degree of correlation between two series of data whose relationship may be described by a straight line. But we have seen that some relations are curvilinear. The method of computing the index of correlation and the index of determination is somewhat different from the methods used to compute the coefficient of correlation and the coefficient of determination. For purposes of illustration the data on per cent of land used for business purposes by census tracts in Indianapolis and the felon rates by census tracts will be used. Referring back to Figure LVIII, it is clear that the relation

TABLE LXXII

Computation of Group Averages to Indicate the Form of the Regression Curve—Crime Data

Per Cent of Land	Felon Rate	Per Cent of Land	Felon Rate	Per Cent of Land	Felon Rate	Per Cent of Land	Felon Rate
0–9.9		10–19.9		20–29.9		30–39.9	
X	Y	X	Y	X	Y	X	Y
4.6	2.7	13.3	5.1	20.0	4.9	33.9	2.8
7.1	2.6	10.6	1.3	23.3	4.0	34.0	3.3
6.8	1.7	16.3	5.7	23.1	13.8		
7.5	2.9	16.6	3.6	22.0	4.4		
5.1	2.1	15.9	1.1	24.3	7.9		
5.3	1.6						
9.8	4.2						
9.0	2.5						
Total 55.2	20.3	72.7	16.8	112.7	38.5	67.9	6.1
Mean 6.9	2.5	14.5	3.4	22.5	7.7	34.0	3.1

[10] See Ezekiel, *op. cit.*, pp. 117, 118.

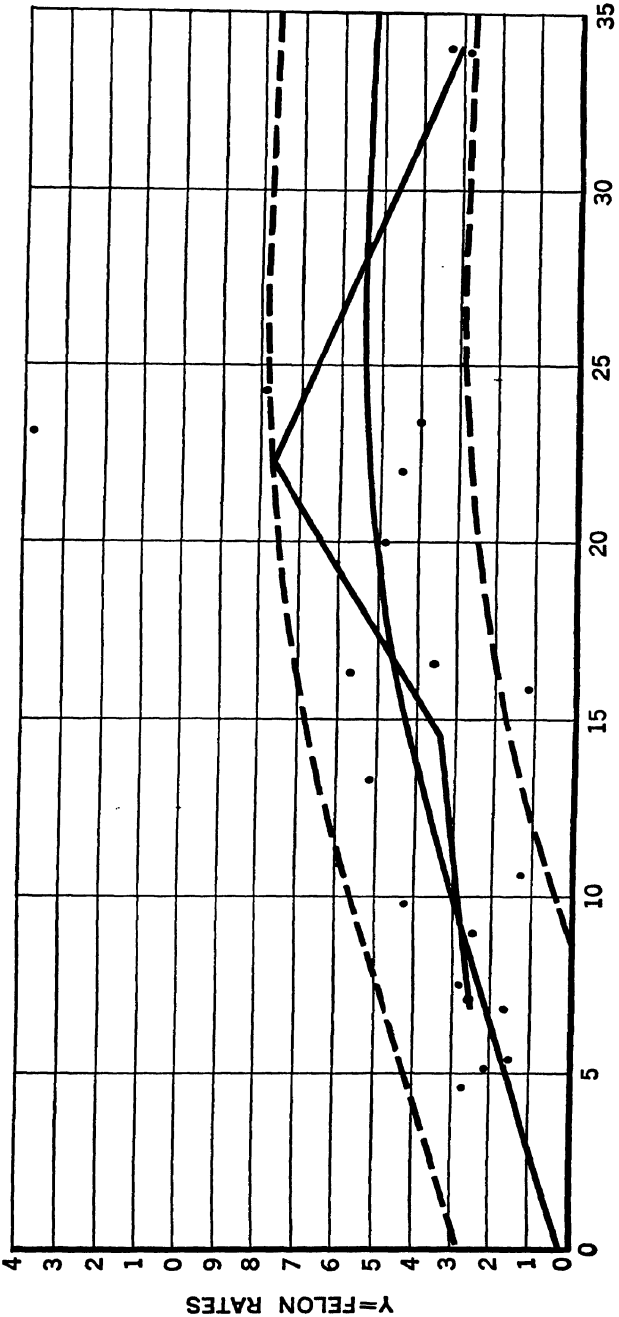


FIGURE LX.—SCATTERGRAM WITH LINE OF MEANS AND FREEHAND CURVE SUPERIMPOSED—CRIME DATA

between these two series is curvilinear. If we had not already determined the fit of a curve to these data, it would not be necessary to fit a freehand curve in order to determine the number of constants for the regression equation. This will be done so that the method may be clear to the student.

If the means of the columns are plotted on a scattergram of the original data, it will be seen that 11 of the dots fall below the line connecting the means, that 8 fall above this line and that one falls on the line. The line of means obviously cannot be represented by a straight line; it is concave downward. Now a smooth curve may be drawn freehand as nearly as possible to fit the data. If it were not for the one extremely high felon rate, the freehand curve would be more concave than it is. Figure LX presents the data. An examination of Figure LVII suggests that the freehand curve approaches nearest to the curve (concave downward) whose equation is $Y = a + b \log X$. These are the same data to which a logarithmic curve was fitted above.[11] The former curve was fitted

TABLE LXXIII

COMPUTATION OF QUANTITIES FOR THE RESIDUALS AND THE STANDARD DEVIATION FOR CURVILINEAR CORRELATION—CRIME DATA

Per Cent of Land X	Felon Rate Y	Felon Rate Estimated from Curve Y'	$Y - Y'$ (z)	$(z)^2$	Y^2
4.6	2.7	1.5	1.2	1.44	7.29
7.1	2.6	2.2	.4	.16	6.76
6.8	1.7	2.1	− .4	.16	2.89
7.5	2.9	2.3	.6	.36	8.41
5.1	2.1	1.7	.4	.16	4.41
5.3	1.6	1.7	− .1	.01	2.56
9.8	4.2	2.8	1.4	1.96	17.64
9.0	2.5	2.7	− .2	.04	6.25
13.5	5.1	3.7	1.4	1.96	26.01
10.6	1.3	3.0	−1.7	2.89	1.69
16.3	5.7	4.2	1.5	2.25	32.49
16.6	3.6	4.3	− .7	.49	12.96
15.9	1.1	4.2	−3.1	9.61	1.21
20.0	4.9	4.8	.1	.01	24.01
23.3	4.0	5.2	−1.2	1.44	16.00
23.1	13.8	5.2	8.6	73.96	190.44
22.0	7.9	5.1	2.8	7.84	62.41
24.3	7.9	5.2	2.7	7.29	62.41
33.9	2.8	5.4	−2.6	6.76	7.84
34.0	3.3	5.3	−2.0	4.00	10.89
Totals 308.7	81.7	72.6	+9.1	122.79	504.57

[11] See pp. 291-293.

mathematically, but it is common practice in computing curvilinear correlation to use the freehand curve from which to read off the values of Y corresponding to values of X on the graph. This will be done in the problem here. The estimated values of Y lie on the smooth logarithmic curve which was drawn freehand. In the computation of the index of correlation and the index of determination logarithms are not used; the actual and estimated data are used. It is necessary to guess the equation of the curve of best fit in order to know how many constants will enter into the equation, because this fact is used in certain parts of the procedure. Table LXXIII shows the process of computing the index of correlation. The comparison of the sum of the Y values with the sum of the Y' values shows the margin of error made in drawing the freehand curve. If they were the same, the sum of the differences, z, would be 0, but instead it is 9.1, the mean of which is .46. Since the sum of the z values is a plus quantity, the mean of these values indicates that the freehand curve should be shifted up .46 units on the Y scale. When the freehand curve is used for estimating Y values, the regression equation may be written:

$$Y = k + f(X)$$

in which k is the constant corresponding to a in the general equation for the curve. This constant is the mean of the sum of the differences between Y and Y', which in our problem is .46. The regression equation may then be written:

in which $f(X)$ may be read "factor of X." To estimate a Y value, then, and include the correction for the error made in drawing the freehand curve, we simply substitute any given value of Y as indicated by a point on the freehand curve and add to it .46.

The totals of the columns in Table LXXIII give the quantities necessary to determine the degree of correlation between per cent of land used for business purposes and the felon rate. The necessary standard deviations may be obtained from the following formulas:

$$\sigma_y = \sqrt{\frac{\Sigma Y^2 - n(M_y)^2}{n}}$$

$$\sigma_z = \sqrt{\frac{\Sigma(z)^2 - n(M_z)^2}{n}}$$

Substituting the appropriate values in these equations and solving we have:

20

$$= \sqrt{\overline{122.79 - 4.23}}$$

From the following formula the index of correlation, corrected for the number of observations, may be determined:

$$\bar{\rho}_{yx}{}^2 = 1 - \left(\frac{\sigma_z{}^2}{\sigma_y{}^2}\right)\left(\frac{n-1}{n-m}\right)$$

Substituting the appropriate values in the equation, we have:

$$\bar{\rho}_{yx}{}^2 = 1 - \left(\frac{5.9049}{10.0489}\right)\left(\frac{20-1}{20-2}\right) = .3$$

$$\bar{\rho}_{yx} = .617$$

The symbol m in the formula refers to the number of constants in the equation of the curve of best fit; in this case it is the guessed logarithmic curve drawn freehand. The index of correlation is found to be .617, and, since the index of determination is simply the square of the index of correlation, the index of determination is .381. The latter represents the per cent of variance in Y which is also present in X; in other words it accounts for 38.1 per cent of the factors entering into the determination of Y.

The same method can be used in working out curvilinear correlation for data in which the curve of best fit is some other logarithmic curve, a hyperbola or a parabola. The constants and the standard deviations to be found would be the same, though in the case of some curves there will be three or more constants instead of two.

It remains to compute the standard error of estimate. This is determined by the formula:

$$_{y.f(x)} = \frac{}{n - m}$$

Substituting the appropriate values in this equation, we have:

$$\bar{S}^2{}_{y.f(x)} \qquad \frac{}{20 - 2}$$

$$\bar{S}_{y.f(x)} = 2.56$$

TABLE LXXIV

Correlation of the Sex Ratio and the Marriage of Women[1]

Sex Ratio—Males per 100 Females	Percentage of Women 25 Years of Age and over Who are Married—X												F_y	d_y	$\Sigma d_x F$	$d_y(\Sigma d_x F)$	$d_y F_y$	$d_y^2 F_y$
Y	44–47	48–51	52–55	56–59	60–63	64–67	68–71	72–75	76–79	80–83	84–87	88–91	(1)	(2)	(3)	(4)	(5)	(6)
60– 68	1												1	−4	−5	20	−4	16
69– 77			2										2	−3	−6	18	−6	18
78– 86		1	1	2	2	1	1						8	−2	−12	24	−16	32
87– 95				5	18	17		1					41	−1	−26	26	−41	41
96–104				1	5	30	18	6					60	0	23	0	0	0
105–113						3	6	9	1				19	1	27	27	19	19
114–122						1	7	10					18	2	27	54	36	72
123–131					1		2	7	1				11	3	18	54	33	99
132–140								2		1			3	4	8	32	12	48
141–149									2	1			3	5	10	50	15	75
150–158									1				1	6	3	18	6	36
159–167											1		1	7	5	35	7	49
168–176													0	8	0	0	0	0
177–185												1	1	9	6	54	9	81
186–194											1		1	10	5	50	10	100
(1) F_x	1	1	3	8	26	52	34	35	5	2	2	1	170	Σ	83	462	80	686
(2) d_x	−5	−4	−3	−2	−1	0	1	2	3	4	5	6	Σ					
(3) $\Sigma d_y F$	−4	−2	−8	−9	−19	−14	24	57	20	9	17	9	80					
(4) $d_x(\Sigma d_y F)$	20	8	24	18	19	0	24	114	60	36	85	54	462					
(5) $d_x F_x$	−5	−4	−9	−16	−26	0	34	70	15	8	10	6	83					
(6) $d_x^2 F_x$	25	16	27	32	26	0	34	140	45	32	50	36	463					

[1] Groves, E. R., and Ogburn, W. F., *American Marriage and Family Relationships*, p. 479. New York: Henry Holt & Co., 1928. This correlation table is substantially that used by Ezekiel, *op. cit.*, p. 354. A few changes in subscripts have been made.

The limits of the standard error are represented by the broken lines in Figure LX. The chances are 2 to 1 that estimates of the value of Y by means of the regression equation above will fall within plus or minus the standard error of estimate, 2.56. This is a large standard error, but it would be reduced by more than a third if the one extreme item were eliminated.

Sometimes the data one wishes to use in computing correlation are so numerous that it would be unnecessarily laborious to work out the correlation by exactly the method used in the preceding illustrations, or the data may already be in the form of frequency tables in which case it would be impossible to determine the separate items. It is, therefore, desirable to have a method of computing the degree of correlation from grouped data. A method for doing this from data whose relationship is indicated by a straight line will now be described.

Data for illustrating the group method of correlation will be taken from some material collected by Professor W. F. Ogburn concerning marriage and the sex ratio. The data are presented in Table LXXIV in the form of a correlation table above.

The symbols have the following meaning:

X = percentage of women 25 years of age and over who are married
Y = sex ratio—males per 100 females
ΣF_x = sum of the frequencies in the columns
ΣF_y = sum of the frequencies in the rows
d_x = step-deviations from assumed mean of X
d_y = step-deviations from assumed mean of Y
$\Sigma d_x F$ = algebraic sum of the x-deviations times the frequencies
$\Sigma d_y F$ = algebraic sum of the y-deviations times the frequencies
$d_y(\Sigma d_x F)$ = product of columns (2) and (3)
$d_x(\Sigma d_y F)$ = product of rows (2) and (3)
$d_x F_x$ = x-deviations times the sum of the frequencies in each column
$d_y F_y$ = y-deviations times the sum of the frequencies in each row
$d_x^2 F_x$ = x-deviations squared times the frequencies in each column
$d_y^2 F_y$ = y-deviations squared times the frequencies in each row

Before computing the coefficient of correlation and the regression equation, certain correction factors must be computed for the deviations from the mean. The corrections to be made are for $\Sigma d_y^2 F_y$, c_y; for $\Sigma d_x^2 F_x$, c_x; and for $\Sigma d_y d_x F$, c_x. The corrected quantities are found in the following manner:

Let $\Sigma y^2 = \Sigma d_y^2 F_y - (\Sigma d_y F) \frac{\Sigma d_y F^{12}}{F}$, corrected y-squares

$\frac{\Sigma d_x F^{12}}{F}$, corrected x-squares

2
-, corrected yx-products

Substituting in these equations the values appearing in Table LXXIV, we have the following results:

o-
648.4
170

$$= 463 - (83)\frac{83}{170} = 422.3$$

$$= 462 - (80)\frac{83}{170} = 422.8$$

The correction for the yx-products is for the regression of Y on X, that is, for the estimation of values of Y from known values of X. If it were desired to estimate values of X from known values of Y, then the correction factor to be used would be c_y to obtain the corrected xy-products. But ordinarily we are concerned only with estimating values of the dependent variable from known values of the independent variable. The corrected values, shown above, are now substituted in the formula for computing the coefficient of correlation:

$$r_{yx} = \Sigma yx$$

$$+422.8$$

$$= +.808$$

The coefficient is quite high, which means that the correlation between the percentage of females 25 years of age and over and the sex ratio is close and positive. The regression equation will now be determined, and the first step is to compute the constants a and b:

$$l \; = \; \frac{\Sigma yx}{}$$

$$= .652 \text{ intervals}$$

[12] The quantities subtracted are the product of the sum of the products of the step-deviations times the frequencies and the mean deviation of each item from the assumed mean group in intervals.

To reduce b_{yx} to terms of scale units compute the ratio of the class-interval of Y to the class-interval of X, as follows:

Multiplying .652 by this figure, we get 1.467 scale units for the value of b_{yx}.

$$\begin{aligned} a &= M_y - b_{yx}M_x \\ &= 104.7 - (1.467)(67.9) \\ &= 5.1 \end{aligned}$$

The regression equation is then:

$$Y = 5.1 + 1.467X$$

Using the same formula as previously used for the correction of the coefficient of correlation for the number of items, we have:

$$\begin{aligned} \bar{r}_{yx}^2 &= 1 - (1 - r^2) \quad n - m \\ &= 1 - (1 - .6529)\,\frac{170 - 1}{170 - 2} \\ &= .6502 \\ \bar{r}_{yx} &= .806 \end{aligned}$$

When the product-moment method of correlation is used, it is customary to write the coefficient plus or minus the probable error. The probable error of the coefficient of correlation above is computed below:

$$\frac{1 - .6496}{13.04}$$

$$= \pm .018$$

The coefficient may then be written:

$$\bar{r}_{yx} = +.806 \pm .018$$

It is usually held that, if the coefficient is 5 or 6 times the size of its probable error, or still greater, it is significant. Since our coefficient is many times greater than the probable error, we may conclude that it is significant.

If the regression equation is used for estimating future values of Y, the estimates should be accompanied by the standard error of estimate. For this purpose, we may use the following formula:

The standard deviation of Y, computed from $\Sigma d_y^2 F_y$, corrected, is 17.4. Substituting in the formula, we have:

$$S_y = 17.4\sqrt{1 - .6496}$$
$$= 10.3$$

It should be noted that, if the product-moment method is used for computing the coefficient of correlation, it is not necessary to use the ordinary equation for regression. An alternative equation is available and is given below:

$$Y - M_y = \bar{r}\frac{\sigma_y}{\sigma_x}(X - M_x)$$

Much of the arithmetic involved in this equation has already been done in the process of deriving the constants, a and b. This equation has no special merit, and the equation, $Y = a + bX$, is in more general use.

4. EXERCISES

1. Below are given two tables. Experiment with different kinds of curves and decide which is the best fit. Compute the equation of the curve in each case:

TABLE LXXV

DIVORCED PERSONS PER 1,000 MALES AND PER 1,000 FEMALES OVER 15 YEARS OF AGE IN CERTAIN CENSUS TRACTS OF INDIANAPOLIS

Divorced Persons Male	Divorced Persons Female
2.2	6.9
2.2	3.2
2.8	13.7
4.9	9.1
5.8	9.4
6.2	14.3
6.4	13.4
6.5	7.7
7.2	5.4
7.2	15.5
8.0	13.0
8.7	15.2
9.1	12.9
9.9	10.5
11.7	27.9
12.7	25.8
14.8	28.1
17.7	25.7
18.8	16.1
19.5	21.9

TABLE LXXVI

AMOUNT OF RELIEF PER RELIEF CASE AND AMOUNT OF RELIEF PER ALLOWANCE CASE IN 20 RELIEF AGENCIES, SEPTEMBER, 1931[1]

Relief per Relief Case	Relief per Allowance Case
$11	$27
15	40
18	24
18	67
19	30
20	24
20	26
25	64
25	32
26	39
27	39
28	32
29	54
31	38
33	51
35	38
37	41
40	47
44	55
48	53

[1] *Monthly Reports*, Department of Statistics, Russell Sage Foundation.

NOTE: A relief case is any case which receives financial assistance from a social agency, but an allowance case is one for which a long-time plan has been made and usually contemplates a large expenditure of funds. Allowance cases usually constitute a small percentage of the total relief case load.

2. The following table gives the number of police per 1,000 population and the number of serious crimes committed per 1,000 population in the month of October, 1931, in 30 cities of 250,000 or more:

TABLE LXXVII

POLICE PER 1,000 POPULATION AND CRIMES PER 1,000 POPULATION IN 30 CITIES, OCTOBER, 1931[1]

City	Police per 1,000 Population	Crimes per 1,000 Population
Akron, O.	.8	1.3
Birmingham, Ala.	1.0	1.8
Dallas, Tex.	1.1	1.5
Columbus, O.	1.2	2.7
Houston, Tex.	1.2	2.8
Minneapolis, Minn.	1.2	1.0
St. Paul, Minn.	1.3	1.1
Oakland, Cal.	1.4	1.8

TABLE LXXVII—(*Continued*)

City	Police per 1,000 Population	Crimes per 1,000 Population
Portland, Ore.	1.5	3.4
Denver, Colo.	1.5	2.2
Cincinnati, O.	1.5	1.9
Toledo, O.	1.5	2.9
Indianapolis, Ind.	1.6	2.8
Louisville, Ky.	1.6	2.0
Rochester, N. Y.	1.6	.9
Kansas City, Mo.	1.7	1.3
Cleveland, O.	1.7	2.0
New Orleans, La.	1.9	.8
Chicago, Ill.	2.0	2.7
Milwaukee, Wis.	2.0	1.0
Buffalo, N. Y.	2.2	.7
San Francisco, Cal.	2.2	2.2
Baltimore, Md.	2.3	1.3
Providence, R. I.	2.4	1.5
Detroit, Mich.	2.6	1.8
Philadelphia, Pa.	2.8	.6
St. Louis, Mo.	2.8	1.8
Washington, D. C.	2.9	3.2
Boston, Mass.	3.3	1.5
Jersey City, N. J.	3.6	.4

[1] *Uniform Crime Reports*, United States Department of Justice, Vol. II, No. 10.

(a) Fit a curve to the data in Table LXXVII.

(b) How important is the relation between police protection and number of crimes committed? Determine this by computing the degree of correlation which exists between the two series of data. Also compute the regression equation and the standard error of estimate.

Table LXXVIII gives the Index of Educational Interest (i.e., the school attendance rate 7 to 13 years of age) and the per cent illiterate in the population 21 years of age or over in 36 Texas Counties in 1920:

TABLE LXXVIII

Index of Educational Interest and Index of Illiteracy, 36 Texas Counties, 1920[1]

County	Index of Educational Interest	Index of Illiteracy
Carson	94.1	1.7
Camp	93.7	14.0
Angelina	93.1	6.9
Cass	91.9	13.1
Bosque	91.2	4.0

TABLE LXXVIII—(*Continued*)

County	Index of Educational Interest	Index of Illiteracy
Armstrong	90.9	1.2
Bell	90.1	4.4
Cherokee	89.5	10.6
Brown	89.0	2.0
Childress	87.8	1.8
Clay	87.2	1.9
Brazoria	86.8	13.3
Chambers	86.5	10.2
Bowie	86.4	11.1
Burleson	86.4	14.1
Anderson	86.1	9.8
Brazos	85.7	16.7
Burnet	85.5	3.8
Austin	85.2	8.3
Castro	84.4	1.1
Arkansas	84.2	8.0
Archer	83.9	2.0
Briscoe	82.0	1.7
Bexar	81.3	13.3
Calhoun	81.3	7.1
Bastrop	81.1	15.9
Blanco	79.7	3.4
Baylor	79.2	1.4
Callahan	78.6	2.9
Bandera	77.5	3.4
Atascosa	65.4	22.8
Bee	63.2	24.0
Cameron	61.8	33.1
Brewster	57.3	31.2
Caldwell	55.0	26.2
Brooks	45.4	34.8

[1] Ross, Frank A., *School Attendance in the United States, 1920*, p. 210.

(a) Fit a curve to the data in this table.
(b) Determine the degree of correlation, the index or coefficient of determination, the regression equation, and the standard error of estimate.
(c) Show graphically the regression curve and the limits of and the standard error of estimate.

4. Table LXXIX gives data for computing the correlation between the sex ratio in the population and the percentage of women 25 years of age or over who are married. The method for grouped data is required:
   (a) Compute the coefficient of correlation for the data in this table.
   (b) Determine the regression equation for the dependent variable and the standard error of estimate.

TABLE LXXIX

THE NUMBER OF MALES PER 100 FEMALES AND THE PER CENT OF WOMEN MARRIED IN 170 CITIES[1]

Per Cent of Women Married—Y	Sex Ratio—Males per 100 Females—X: 60–68	69–77	78–86	87–95	96–104	105–113	114–122	123–131	132–140	141–149	150–158	159–167	168–176	177–185	186–194	Total
88–91														1		1
84–87												1			1	2
80–83									1	1						2
76–79						1		1		2	1					5
72–75				1	6	9	10	7	2							35
68–71			1		18	6	7	2								34
64–67			1	17	30	3	1									52
60–63			2	18	5			1								26
56–59			2	5	1											8
52–55		2	1													3
48–51			1													1
44–47	1															1
Total	1	2	8	41	60	19	18	11	3	3	1	1		1	1	170

[1] Groves, E. R., and Ogburn, W. F., *American Marriage and Family Relationships*, p. 479. New York: Henry Holt & Co., 1928.

(c) Show the regression line and the standard error of estimate graphically.

5. In order that the student may gain practice in thinking in terms of functional relations, let each student obtain data:
   (a) Which show linearity and are ungrouped.
   (b) Which show curvilinearity and are ungrouped.
   (c) Which show linearity and are grouped.
   (d) In each case compute the degree of correlation, the regression equation, the coefficient or index of determination, and the standard error of estimate.
   (e) In each case present graphically the regression equation and the standard error of estimate.

5. REFERENCES

Chaddock, Robert E., *Principles and Methods of Statistics*, Chap. XII.

Ezekiel, Mordecai, *Methods of Correlation Analysis*, Chaps. 3-9.

Mills, Frederick C., *Statistical Methods*, Chaps. X, XII, XIII.

Thurstone, L. L., *The Fundamentals of Statistics*, Chaps. 22-24.

CHAPTER XII

The Theory of Probability

1. INTRODUCTION

STATISTICS is concerned with chance variations, or probabilities. It is, therefore, not surprising that the first persons who became seriously interested in the theory of probability were gamblers. As early as the fifteenth century various European mathematicians were asked by gamblers to calculate the probabilities of winning in games of chance. The names of Pascal, Fermat, and Leibnitz appear among those consulted by gamblers. The first scientific treatise on the subject was written in Latin; it was published November 12, 1733, by De Moivre. It approached the problem by the method of binomial expansion and was intended to be a guide to gamblers. In the early part of the eighteenth century astronomers became interested in probability, and the number of mathematicians interested in it increased. Among those who made important contributions to the subject were Laplace and Gauss. Serious interest in the theory of probability, then, had an empirical origin. Since it began to attract wide attention among mathematicians much work has been done on it, but in books on statistics the chief interest is still empirical. Natural and social phenomena seem to occur or vary according to the laws of probability; hence, every step in social statistics involves the theory of probability.[1]

In reading the preceding chapters and working out the problems in connection with methods described, the student must have been aware that he was dealing with a chance distribution of measurements or counts. At all times it has been clear that a statistical result was a "probable result" within certain limits of variability.

[1] For a good summary of the history of the theory of probability, see Walker, Helen M., *Studies in the History of Statistical Method,* Chap. II. Baltimore: Williams & Wilkins, 1929.

The measures of dispersion—quartile deviation, average deviation, and standard deviation—are frank admissions that an average is only the most likely value and that, in fact, any sample of data taken from a universe will show scatter above and below the average. In a distribution which approaches the symmetrical bell-shaped form 50 per cent of the values will fall between the median minus and plus once the quartile deviation; that is, the chances, or probabilities, are even that any value selected at random will be neither less nor greater than the median minus and plus once the quartile deviation. In such a distribution 57.5 per cent of the values will fall between the average minus and plus once the average deviation. The corresponding limits of the standard deviation from the mean include 68.26 per cent of the values. Here we are speaking of chance, or probability, but it is chance with reference to the specific data in hand and not with reference to the universe of data from which the sample was drawn. Normal probability is a concept derived from the distribution of all the values in the universe or upon a sample indefinitely large. Any particular sample must be referred to the normal distribution of the universe of data as its standard of accuracy. The standard error of estimate of a regression equation is likewise a measure of the chances of occurrence of an event. It involves the theory of probability. Instead of saying that we can estimate the value of Y from a known value of X, we say that we can estimate the value of Y within certain limits of variability, or within the limits of its standard error. Obviously the smaller the standard error, the greater the reliability of estimates.

The term "error" in statistics does not refer to mistakes. Mistakes arise from hasty or careless work or from inaccurate perception. To err means to wander from a path or a norm. In every universe of data there is a central value about which it is normal for the individual measures to err or wander. Errors, in this sense, can be determined mathematically. The probability of the occurrence of an event of a certain magnitude is the chance out of a finite number of possible events that the particular event will occur.

2. ELEMENTARY ILLUSTRATIONS OF PROBABILITY

If a coin is tossed, one of two things may happen: the tail will turn up or the head will turn up. The chances are even that the coin will fall tail up or head up. How may this fact be expressed in symbols? Let p represent success, q represent failure, and n

represent the total ways in which the event may occur. A head will be represented by a and a tail by b. Then, if a head may be regarded as success and a tail as failure, the chances of a head falling may be expressed thus:

and the chances of failure are:

Or, $$p = \frac{1}{2}$$

And $$q = \frac{1}{2}$$

But suppose that instead of there being only 2 possible events, there are 52, as there would be in drawing a particular card from a complete deck of cards. The chance of drawing a jack of hearts from a deck of cards is:

$$p = \frac{a}{n} = \frac{1}{52}$$

The chance of drawing *any* heart from the deck would be $\frac{1}{4}$, because one-fourth of all the cards are hearts. The probability of an event occurring is the ratio of the event to the total number of possible events.

But suppose there are two alternatives out of a large number of possible events. What would be the chance of drawing either a jack of hearts or an ace of diamonds from a deck of cards? The probability of one or the other of these events happening is the sum of the separate probabilities and may be expressed thus:

$$p = \frac{c}{n} + \frac{d}{n} = \frac{1}{52} + \frac{1}{52} = \frac{1}{26}$$

If we think of the drawing of these two cards as two separate withdrawals, we have a compound event. Neither is dependent upon the other, and two cards are to be drawn. Under such circumstances the chance of drawing a jack of hearts and an ace of diamonds is the product of the probabilities:

$$p = \frac{c}{n} \times \frac{d}{n} = \frac{1}{52} \times \frac{1}{52} = \frac{1}{2704}$$

What has been indicated in simple terms can be expressed in general terms as the expansion of a binomial. Since the tossing of a coin is about as nearly uncontrolled by any factors outside of gravitation as any event is likely to be, we shall continue the coin illustration. If we toss two coins four times, there are four possible combinations of heads and tails:

1 head, 1 tail
2 heads
2 tails
1 tail, 1 head

What are the chances of securing two heads, no heads, and one head? The chances of securing two heads are $\frac{1}{4}$; of securing one head, $\frac{2}{4}$; of securing no heads, $\frac{1}{4}$. Similarly, the chances of securing a certain number of heads and tails could be determined if 5 coins or 10 coins were used. This is a problem in binomial expansion. Using p and q with the same meaning as above, the following binomial holds for 2 coins:

Or, since the chances of at least one head or at least one tail are $\frac{1}{2}$, we may express it with the numbers thus:

4 4 4

The number of coins determines the power of the binomial. If we should use 5 coins, the binomial would be:

$$(p + q)^5 = p^5 + 5p^4q -$$

and we should have the following if numbers are used:

32 · 32 · 32 · 32 · 32 · 32

If we should throw the 5 coins 100 times, the number of the above combinations would be 100 times the numerator of each term in the expanded binomial, or 100, 500, 1000, 1000, 500, and 100, respectively. That is the theoretical distribution which would result. If it were actually done, the numerators of the terms would vary some from these even quantities. However, if the coins were thrown 10,000 times, the chances of a distribution proportionate

to the numerators of the terms in the expanded binomial would be good. The larger the number of throws, the more closely to the theoretical distribution the result is likely to be.

An experiment to determine the nearness of actual successes to theoretical successes was made by Mr. W. F. R. Weldon. He took 12 dice and threw them 4,096 times. A throw which turned up 4, 5, or 6 points was regarded as success, and a throw which turned up 1, 2, or 3 was regarded as failure. This number of throws is sufficiently large to approach the theoretical distribution of successes. Table LXXX gives the number of successes for each throw and the frequencies:

TABLE LXXX

COMPARISON OF ACTUAL AND THEORETICAL SUCCESS FREQUENCIES IN 4,096 THROWS OF 12 DICE [1]

Number of Successes	Frequency, Actual	Frequency, Theoretical
0	0	1
1	7	12
2	60	66
3	198	220
4	430	495
5	731	792
6	948	924
7	847	792
8	536	495
9	257	220
10	71	66
11	11	12
12	0	1

[1] For the actual frequencies, see Yule, U. G., *op. cit.*, p. 258, or the *Encyclopedia Britannica*, 11th ed., Vol. XXII, p. 394, article by F. Y. Edgeworth.

In any particular throw of the 12 dice it is possible to have 0 successes or as many as 12 successes. The theoretical frequencies represent the expansion of $(p + q)^{12}$. An examination of the theoretical frequencies will reveal the fact that the distribution is perfectly symmetrical. The actual frequencies approach the theoretical proportions, but they vary from them slightly at every point. In order to show more clearly the relation between the two distributions, they are presented graphically in Figure LXI: The two curves are quite similar. Obviously, if enough throws of the dice were made, the empirical curve would approach closer and closer to the form of the theoretical curve based upon the expansion of the binomial. Where either of two events may happen

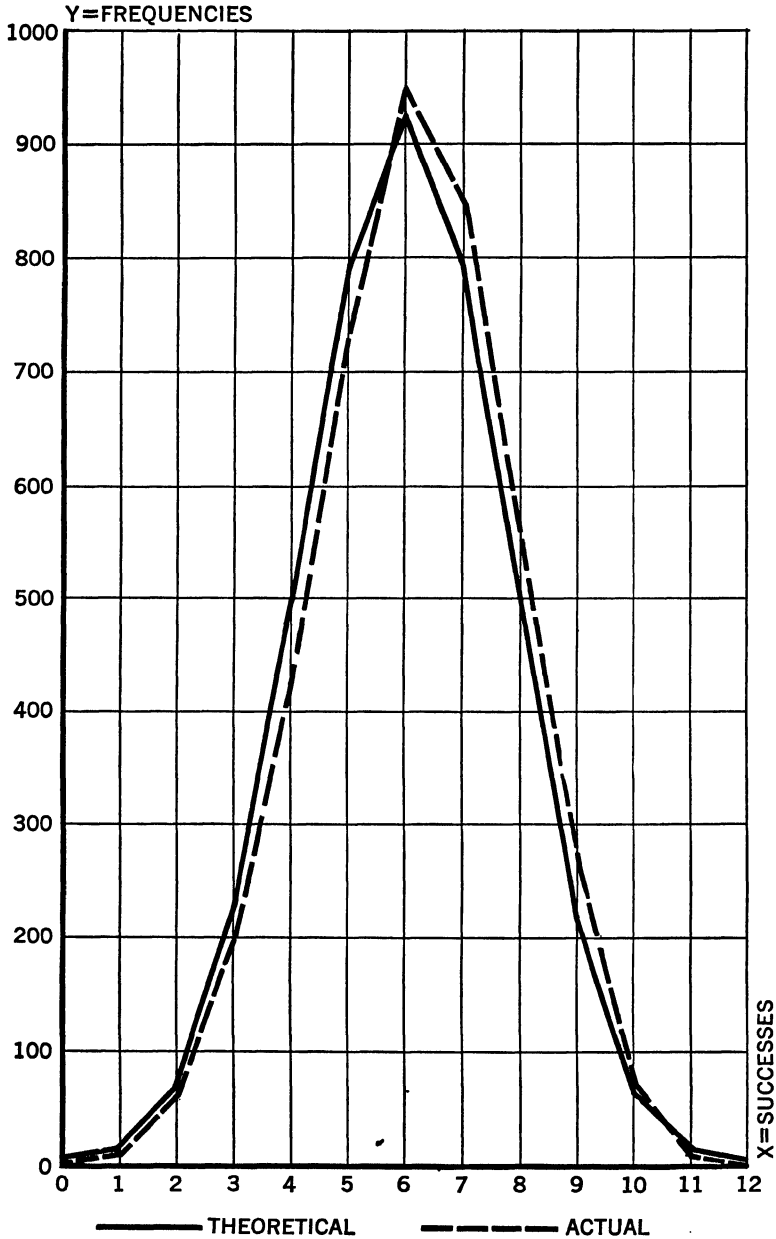


FIGURE LXI.—NUMBER OF SUCCESSES (X) AND ACTUAL AND THEORETICAL FREQUENCIES (Y) IN 4,096 THROWS OF 12 DICE

and where no forces except chance operate, the law which describes their occurrence is the normal curve, as shown in Figure LXII.

The theoretical mean is $M = 6.0$, and the theoretical standard deviation is $\sigma_{12} = 1.732$. The actual mean is $M = 6.139$, and the actual standard deviation is $\sigma_{12} = 1.712$. It is very simple to determine the mean and the standard deviation of the theoretical distribution. The formulas are as follows:

$$M_{12} = np$$
$$\sigma_{12} = \sqrt{npq}$$

in which n is the number of dice, and p and q have the same meaning as above. The same formula would be used for any number of dice which might be used. This number determines the power of the binomial, and the number of terms in the expanded binomial will be one more than the power to which the binomial is raised.

The principal value of the theoretical curve lies in the fact that it provides a basis of generalization. Any sample taken from a universe of data which theoretically are distributed according to the normal curve will vary more or less from the smooth curve. That variance is a measure of the atypicality of the sample; there were chance fluctuations in the selections of the sample, or there was a bias which led to error. As previously suggested, this theoretical curve is variously known as the normal curve of error, the bell-shaped curve, the perfectly symmetrical curve, or the Gaussian curve.

3. THE NORMAL CURVE OF ERROR

Some further explanation of the normal curve of error is desirable in order to show its uses in practical statistical work. The concept of errors will be clearer if Figure LXII is examined. The diagram is made on the basis of rectangular coördinates to emphasize the nature of statistical error—not statistical mistakes. In Figure LXII, YO indicates the value, that is, the mean, at which the largest number of frequencies occur in the normal distribution, such as the theoretical distribution of successes in the coin throwing experiment. It may be referred to as the *zero ordinate*. Any X-value besides the mean will have less frequencies than the mean value of X. There are as many values of X less than the mean as there are values of X greater than the mean. Values of X to the right of O are plus values, and values of X

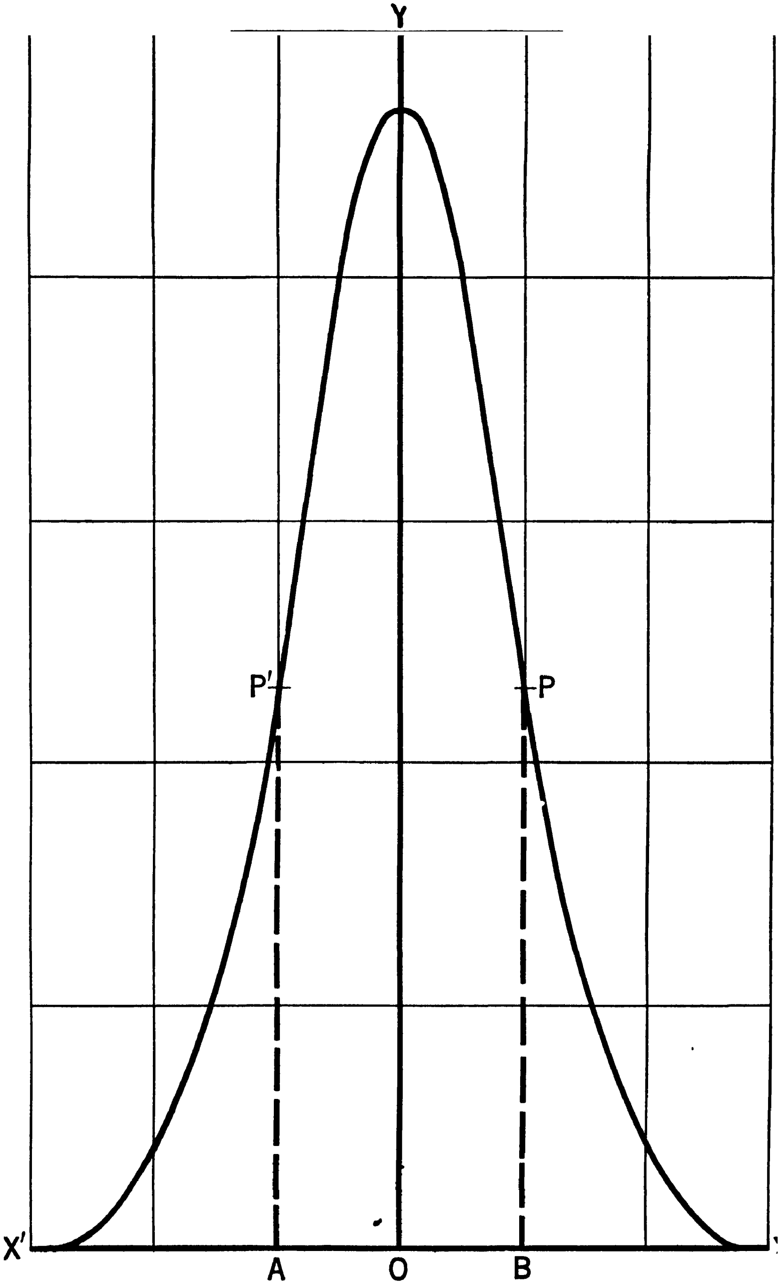


FIGURE LXII.—THE NORMAL CURVE OF ERROR

to the left of O are minus values, with respect to the mean. In the dice throwing experiment the most likely result of any throw is 6 successes. If 6 success values are not thrown, the chances are even that the number of successes will be above or below the mean. We could obtain a similar distribution of data if we measured the heights of all schoolboys 12 years of age in a large city; their heights would be distributed approximately in the form of the normal curve—of the measures deviating from the mean height, half would likely be above the mean and half below. Any small sample of boys might reveal a height distribution varying considerably from the normal curve. Graphic comparison of the curve of the sample with the theoretical curve would indicate roughly the degree of agreement.

We can, however, determine the degree of similarity between a given and a normal frequency distribution by the method of moments. The procedure for fitting a theoretical curve will be described later, but at this point the computation of the moments of a frequency distribution will be illustrated. The following table gives the data required and the first arithmetical step:

TABLE LXXXI

COMPUTATION OF VALUES REQUIRED FOR THE DETERMINATION OF MOMENTS--INTELLIGENCE TEST DATA [1]

Class-Interval X (1)	Mid-Point m (2)	Frequency f (3)	Step-Deviations x (4)	fx (5)	$f(x)^2$ (6)	$f(x)^3$ (7)	$f(x)^4$ (8)
50– 59.9	55	11	−5	− 55	275	−1375	6875
60– 69.9	65	59	−4	−236	944	−3776	15104
70– 79.9	75	149	−3	−447	1341	−4023	12069
80– 89.9	85	256	−2	−512	1024	−2048	4096
90– 99.9	95	328	−1	−328	328	− 328	328
100–109.9	105	352	0				
110–119.9	115	249	1	249	249	249	249
120–129.9	125	165	2	330	660	1320	2640
130–139.9	135	68	3	204	612	1836	5508
140–149.9	145	22	4	88	352	1408	5632
150–159.9	155	8	5	40	200	1000	5000
160–169.9	165	2	6	12	72	432	2592
170–179.9	175	2	7	14	98	686	4802
		1671		−641	6155	−4619	64895

[1] Goodenough, Florence L., *Measurement of Intelligence by Drawings*. Yonkers: World Book Co., 1926. See p. 46, Table 8, last column.

The sums of the four columns containing the products of the frequencies and powers of x are known as the moments of the

distribution about an arbitrary origin. The term "moment" is borrowed from mechanics and refers to the force required to produce rotation about a point. The greater the distance of the application of the force from the axis of rotation, the greater the power of the force. In statistics the frequencies of the various class-intervals are regarded as the forces, and the axis of rotation is the arbitrary origin from which the step-deviations are measured.

The moments about the arbitrary origin are computed as follows:

$$\nu_1 = \frac{\Sigma fx}{n} = \frac{-641}{1671} = -.383, \text{ the first moment}$$

$$\cdot\cdot - \frac{\Sigma f(x)^2}{} - \frac{6151}{1671} = 3.683, \text{ the second moment}$$

$$\frac{-4619}{1671} = -2.764, \text{ the third moment}$$

$$n \qquad 1671 \qquad \text{the fourth moment}$$

But it is not the moments about the arbitrary origin which are of most importance: it is the moments about the mean. These are computed in the following manner:

$\pi_1 = 0$, first moment about the mean
$\pi_2 = \nu_2 - \nu_1^2 = 3.536$, second moment about the mean
$\pi_3 = \nu_3 - 3\nu_1\,\nu_2 + 2\nu_1^3 = 1.176$, third moment about the mean
$\pi_4 = \nu_4 - 4\nu_1\,\nu_3 + 6\nu_1^2\,\nu_2 - 3\nu_1^4 = 37.721$, fourth moment about the mean

W. F. Sheppard has shown that, because of the grouping into class-intervals, certain corrections should be made in the second and fourth moments. The corrected moments are as follows:

$$\mu_1 = 0$$
$$\mu_2 = 3.536 - 1/12 = 3.453$$
$$\mu_3 = 1.176$$
$$\mu_4 = 37.721 - 1/2\pi_2 + 7/240 = 35.982$$

From the corrected moments we obtain two other functions which enable us to determine whether or not the distribution is of the type of the normal curve. These are determined as follows:

$$44.174$$

$$= 3.018$$

$$\mu_2{}^4$$

For the normal curve these functions are:

It is, therefore, clear that the distribution of intelligence quotients approaches closely to the type of the normal curve.

It is worth while noting that the standard deviation in intervals of the distribution is equal to the square root of μ_2. Thus:

However, it is desirable to have a method of fitting a theoretical curve to actual data which seem to conform to the normal distribution. This can be easily done by reference to a table of integrals, because the height of any ordinate above or below the zero, or maximum, ordinate bears a definite relation to the height of this maximum ordinate. This ordinate is called the maximum ordinate because it represents the greatest number of frequencies of any ordinate that can be drawn. The most common equation for the normal curve is:

$$y = y_0 e^{\frac{-x^2}{2\sigma^2}}$$

in which y is the particular ordinate desired; y_0 is the maximum ordinate; e is a constant with the value of 2.7182818 (the base of the Napierian logarithms); x is the value of the independent variable for which the ordinate is to be determined; and σ is the standard deviation of the data. The use of this formula is rather complicated. If the maximum ordinate is known, the relative size of other ordinates may be read from a table of integrals, and the computation is then simple. The formula for determining the maximum ordinate is:

$$y_0 = \frac{\quad}{\sigma\sqrt{2\pi}}$$

or,

$$y_0 = \frac{n}{2.5066\sigma}$$

In this formula σ should be expressed in intervals. The heights of other ordinates may be read from Table CXXI in Appendix A. The height of the ordinate is determined by its distance in terms of standard deviation from the mean. For example, if it is desired to know the height of the ordinates $.5\sigma$ above and below the maximum ordinate, we look at Table CXXI and find .5. To the

right in the first column is the number 88250, or it is 88.250 per cent of the height of the maximum ordinate. Knowing the frequencies represented by the maximum ordinate, we take 88.250 per cent of these frequencies, and the result is the frequencies of the ordinates above and below the maximum ordinate at .5σ removed. In a similar manner other ordinates can be computed. For purposes of illustration some intelligence test data will be used. They are given in the following table:

TABLE LXXXII

I.Q.'s of 1,671 Children, Ages 6 to 12

I.Q.	Number of Children
50– 59.9	11
60– 69.9	59
70– 79.9	149
80– 89.9	256
90– 99.9	328
100–109.9	352
110–119.9	249
120–129.9	165
130–139.9	68
140–149.9	22
150–159.9	8
160–169.9	2
170–179.9	2

TABLE LXXXIII

Fractions of Sigma, Ratio of y to y_0, and Theoretical Frequencies for the Normal Curve

Deviations from Mean in Fractions of σ	Normal Curve y/y_0	y
.0	1.0000	355
.2	.9802	348
.4	.9231	328
.6	.8353	296
.8	.7262	258
1.0	.6065	215
1.2	.4868	173
1.4	.3753	133
1.6	.2780	99
1.8	.1979	70
2.0	.1353	48
2.2	.0889	32
2.4	.0561	20
2.6	.0341	12
2.8	.0198	7
3.0	.0111	4
4.0	.0003	0.1
5.0	.0000	0.0

It will be seen that these data are distributed approximately in the form of a normal frequency curve. How close does this distribution approach the normal distribution of I.Q's of the same mean and the same standard deviation? Table LXXXIII gives the theoretical frequencies of the fitted normal curve for the I.Q. data. The symbol y represents any particular theoretical frequency, and y_0 represents the theoretical maximum ordinate. Each of the frequencies below the frequency of the maximum ordinate in this table will appear above and below the maximum ordinate in the complete distribution and in the graph of the curve; that is, we have to use both plus and minus fractions of the standard deviation as measured from the mean ordinate. Figure LXIII shows how actual and theoretical distributions compare (see next page). The curves are quite similar, but they coincide at only a few points. The difference may be explained in either of two ways: (1) the failure to fit may be due to chance variations in the sample, which would be eliminated if a large number of I.Q's were taken; (2) or it may be that I.Q's are not distributed according to the normal curve. This question can be answered, but the distribution of frequencies must be recalculated in terms of the area of the frequency polygon. If the fit of the curve is sufficiently close, it is reasonable to conclude that I.Q's are distributed according to the normal curve and that the fluctuations are due to errors in sampling.

The computation of frequencies in terms of the area of the frequency polygon is somewhat more laborious than their computation in terms of the maximum ordinate, but the test of goodness of fit is in terms of the former. It was indicated in Table LXXXIII that the maximum, or zero, ordinate is unity, or 100.0 per cent. Likewise, the total area of the frequency polygon is regarded as unity. The object of the computations is to determine the proportion of frequencies in the area enclosed by the maximum ordinate and any other ordinate above or below it. After the deviations from the mean in intervals are determined and expressed as fractions of the standard deviation, the proportion of frequencies between the maximum and any other ordinate may be found in Appendix A, Table CXXII. Table LXXXIV shows the method of computing the theoretical distribution of the 1,671 I.Q's.

The value of y is obtained by multiplying 1,671 by the value of y/y_0 for each class-interval. The total of the theoretical distribution is two-tenths more than the total of the actual frequencies. If the ratio of y to y_0 had been carried to one or two more decimal

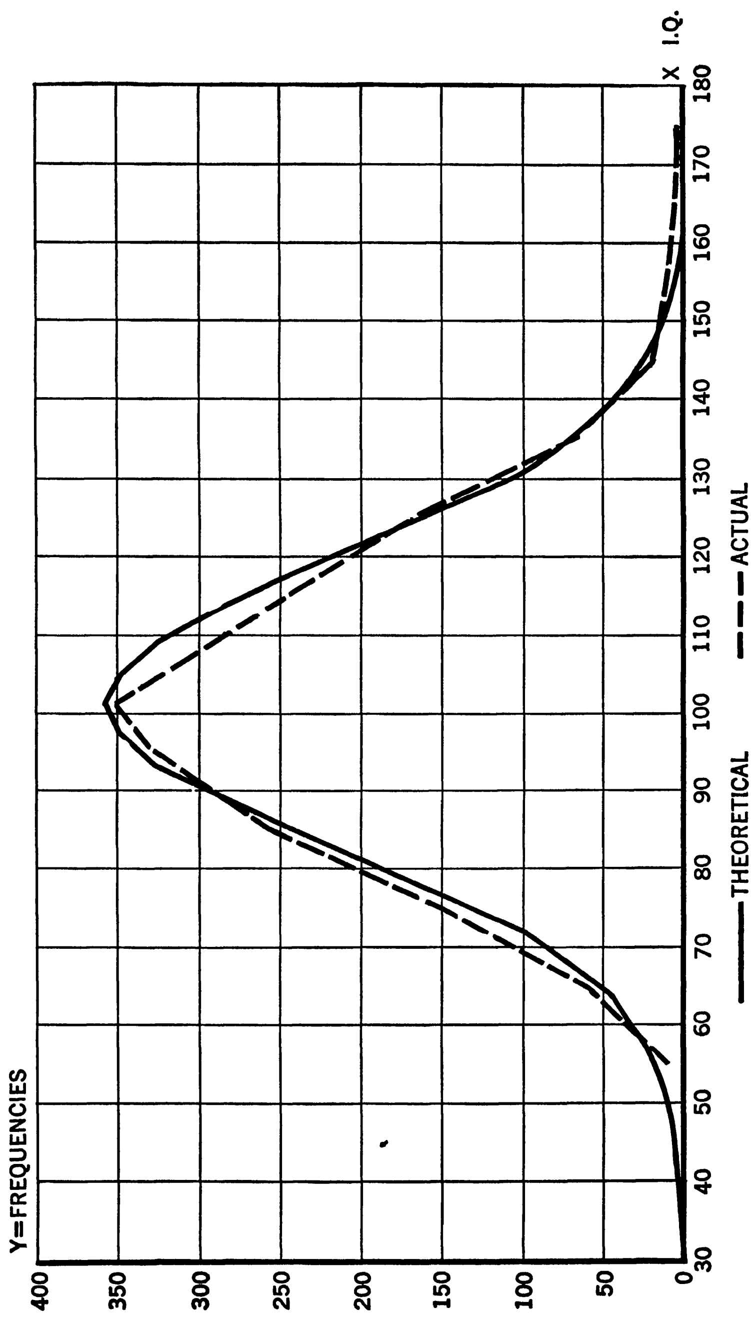


FIGURE LXIII.—NORMAL CURVE DETERMINED FROM ORDINATES EXPRESSED AS FRACTIONAL PARTS OF THE MAXIMUM ORDINATE, COMPARED WITH ACTUAL DATA

TABLE LXXXIV

COMPUTATION OF THEORETICAL FREQUENCIES FOR 1,671 I.Q.'S

Class-Intervals in I.Q.'s X	Class Limits Lower	Class Limits Upper	Deviations from Mean in Intervals x	Deviations from Mean $\div \sigma$ in Intervals x/σ	Proportion of Area between y_0 and Ordinate y/y_0	Cases between y_0 and Ordinate y	$N =$ 1671 f
Below 40					.5000	835.50	1.00
40– 49.9	40		−6.12	−3.26	.4994	834.50	4.52
50– 59.9	50		−5.12	−2.72	.4967	829.98	18.38
60– 69.9	60		−4.12	−2.19	.4857	811.60	48.96
70– 79.9	70		−3.12	−1.71	.4564	762.64	141.38
80– 89.9	80		−2.12	−1.13	.3718	621.28	249.63
90– 99.9	90		−1.12	− .59	.2224	371.63	331.69
100–109.9	100		− .12 } 0	− .06 }	.0239	39.94 }	336.04
		110	.88 }	.46 }	.1772	296.10 }	
110–119.9		120	1.88	1.00	.3413	570.31	274.21
120–129.9		130	2.88	1.53	.4370	730.23	159.92
130–139.9		140	3.88	2.06	.4803	802.58	72.35
140–149.9		150	4.88	2.60	.4953	827.65	25.07
150–159.9		160	5.88	3.13	.4991	834.00	6.35
160–169.9		170	6.88	3.66	.4999	835.33	1.33
Above 170					.5000	835.50	.17
							1671.20

$M = 101.2$ $\sigma = 18.8$, units of I.Q.
$= 1.88$, class-intervals

places, the totals should have been identical. However, this variation of .2 does not materially affect the size of the frequencies. The last column is obtained from the y's: The maximum ordinate is determined by adding 39.94 and 296.10, the frequencies in the two parts of the class-interval, 100-109.9. The frequencies in this class-interval are in two parts, because one part is below the mean and one part is above. The frequency in the class-interval, 90-99.9, is found by subtracting 39.94 from 371.63, and the other frequencies below the mean are found in a similar manner by subtracting from the given y-value below it. The frequency for the class-interval, 110-119.9, is found by subtracting 296.10 from 570.31. The other frequencies above the mean are found by subtracting from the given y-value the y-value immediately above it. The results are given as f in the last column. It should be noted that x, which is in terms of intervals, should be divided by σ in terms of intervals.

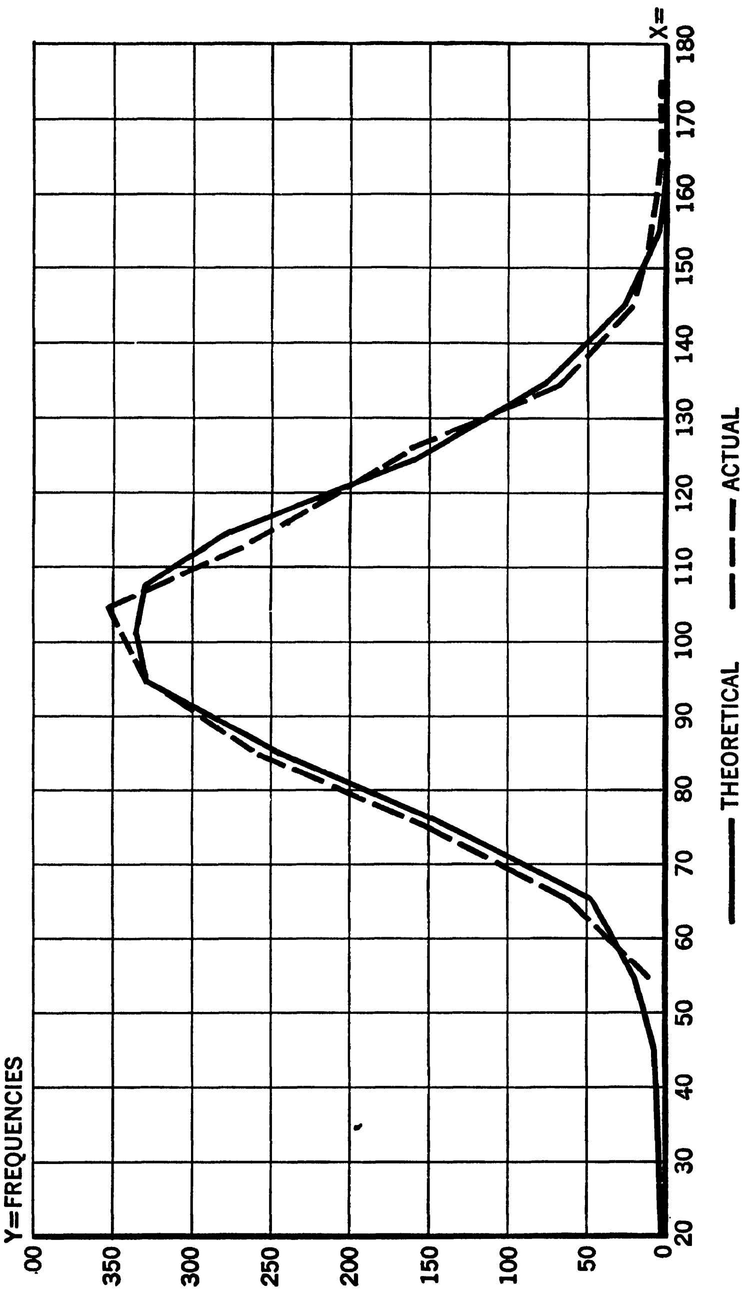


FIGURE LXIV.—NORMAL CURVE DETERMINED FROM RATIO OF Y TO Y_0, COMPARED WITH ACTUAL DATA

Figure LXIV presents the actual and the theoretical frequencies. As in Figure LXIII, it is clear that the normal curve is not an exact fit for the data. The problem now is to determine whether or not it is a sufficiently close fit to justify the conclusion that I.Q's are distributed according to the normal curve of error.

It was pointed out above that the standard deviation may be determined from the formula:

$$\sigma = \sqrt{npq}$$

in which n is the number of events, p the probability of success, and q the probability of failure. In dealing with a frequency distribution the formula has to be altered somewhat, as follows:

in which f is the theoretical frequency at a given point on the X-scale and N is the total number of items. Then,

let $$q = \frac{N-f}{N}$$

and substitute N for n in the general formula, as follows:

N

This is called the standard error of sampling.

We may now set up a table to show the differences between actual and theoretical frequencies.

The absolute differences are not large, but the size of some of the differences relative to the frequencies is fairly large. We shall employ the formula for the standard error of sampling to see the significance of two of the variations. Let us take the first class-interval and the class-interval 110-119.9:

$$\frac{3.38(1671 - 18.38)}{1671}$$

TABLE LXXXV

Differences between Actual and Theoretical Frequencies

Class-Interval m	Actual Frequency f_0	Theoretical Frequency f	Differences f_0-f
55	11	18.38	− 7.38
65	59	48.96	10.04
75	149	141.38	7.62
85	256	249.63	6.38
95	328	331.69	− 3.69
105	352	336.04	15.96
115	249	274.21	−25.21
125	165	159.92	5.08
135	68	72.35	− 4.35
145	22	25.07	− 3.07
155	8	6.35	1.65
165	2	1.33	.67

The standard error of sampling is 4.26. Since the difference between the actual and the theoretical frequency is —7.38, the deviation from the mean is 1.7 times σ_s. If we consult Appendix A, Table CXXII, we find that, when x/σ is 1.7, the proportion of the total area of the frequency polygon included between the maximum ordinate and an ordinate erected at 1.7σ is .4554. The area included between the ordinates erected at 1.7σ above and below the maximum would be equal to 91.08 per cent of the total area. The chances are about 9 out of 100 that a given value will differ from the mean by more than 1.7σ. This is a fairly large deviation. Let us try another class-interval:

The standard error of the sample divides into the difference between the actual and the theoretical frequencies 1.7 times. Referring to Table CXXII in Appendix A, it is seen that $\pm 1.7\sigma$ from the maximum ordinate would include 91.08 per cent of all the frequencies, or, the chances are that about 9 times out of 100 a given value would differ from the mean by more than 1.7σ. This still suggests a rather wide variation, though the standard error might be due to fluctuations of sampling.

If other class-intervals were used for computing the standard error of sampling, we should probably get some variation from the two already computed. Some method is needed by which account may be taken of all the class frequencies. Karl Pearson has developed a method which is known as the Chi-Square Test

of Goodness of Fit. Table LXXXVI gives the data and the computations necessary for determining χ^2. χ^2 is the sum of the squares of the differences between the actual and the theoretical frequencies divided by the theoretical frequencies.

TABLE LXXXVI

COMPUTATION OF χ^2

Class-Intervals X	Actual Frequencies f_0	Theoretical Frequencies f	$f_0 - f$	$\frac{(f_0 - f)^2}{f}$
Below 60	11	23.90	−12.90	6.96
60– 69.9	59	48.96	10.04	2.06
70– 79.9	149	141.38	7.62	.41
80– 89.9	256	249.63	6.35	.16
90– 99.9	328	331.69	− 3.69	.04
100–109.9	352	336.04	15.96	.76
110–119.9	249	274.21	−25.21	2.32
120–129.9	165	159.92	5.08	.16
130–139.9	68	72.35	− 4.35	.26
140–149.9	22	25.07	− 3.07	.38
Above 150	12	7.85	4.15	2.19
	1671	1671.00		15.70

χ^2 is 15.70. From Elderton's table we find that when n', the number of class-intervals, equals 11 and χ^2 equals 15, the probability integral is .132061, and when n' is 11 and χ^2 is 16, the probability integral is .099632.[2] The value of χ^2 in our problem lies between these two values; so it is necessary to interpolate to find the exact value of the probability integral. It proves to be .109461. This means that out of 100 samples of I.Q's, the same size as the one used here, the chances are that about 10.9 would vary farther from the normal curve than the present sample. Two inferences from this fact follow in so far as the present sample is concerned. First, the fact that only about 11 per cent of other samples would vary farther from the theoretical curve than our sample suggests that ours is not a very good one, as samples of I.Q's go, because about 89 per cent of other samples would be nearer to a normal distribution. Second, in view of the fact that the present sample approaches the form of a normal distribution and yet compared with other samples is not a very good one, it seems reasonable to conclude that I.Q's are distributed according to the normal curve and that the theoretical curve fits the distribution.

[2] Pearson, Karl, *Tables for Statisticians and Biometricians*. London: Cambridge University Press, 1924.

4. ESTIMATION OF ERROR IN SAMPLES

The tests of goodness of fit reduce the alternative explanations of the variations of the actual from the theoretical data to two: if the theoretical curve does not closely fit the actual data, the explanation may be that the sample is not representative of the universe of data, or it may be that this universe of data is not distributed according to the theoretical curve selected. Successive samples may be taken and compared with the theoretical distribution. If the standard errors of the samples determined from the formula

$$\sqrt{\frac{f(N-f)}{N}}$$

are not uniformly too large, it is reasonable to assume that an indefinitely large sample selected at random would approach closely to the theoretical distribution. On the other hand, if the standard error of sampling is persistently so large that estimates based upon it would be meaningless, the chances are that the distribution has a form different from the curve selected to represent the data.

There are two measures of reliability in use: the probable error and the standard error. The probable error is based upon the quartile deviation, and the standard error is based upon the standard deviation. Both are equally good as measures of reliability. That is obvious from the fact that there is a constant relation between the two in a distribution which conforms to the normal curve of error. The probable error is .6745 of the standard deviation in a normal distribution. For this reason error computed in terms of one measure may be reduced to terms of the other. However, there are not equally good practical reasons for using the two measures. The standard error is more commonly used, and most of the published tables used in fitting curves to data have been computed in terms of the standard deviation. Hence, it is of practical importance for the student to understand clearly the standard error. Besides measures of error of samples, there are measures of error for other statistical measures. Several of the more common ones will be described.

It was stated above that one may test the representativeness of a sample by taking successive samples and comparing them. This is undoubtedly the best method. But it is laborious and requires a

great deal of time. Frequently it is not practicable. In such cases the standard error of sample means, standard deviations, etc., may be determined from the mean, standard deviation, or other known measure. This shows within what limits these measures for other samples of the same size might be expected to vary.

For example, the standard error of the mean may be determined from this formula:

$$\text{S.E.}_M \qquad \sqrt{N}$$

If we substitute in this formula the appropriate values obtained from the intelligence test problem, we have:

$$\text{S E} \qquad \frac{18.8}{\sqrt{(1671)}}$$

$$\pm .46$$

The chances are 2 to 1 that the mean of any other sample of the same size would not be less than 100.74 or greater than 101.66. That is a small range of fluctuation. The mean should be written 101.2 ± .46. That shows clearly, then, the limits of probable variation. If it were desired to use the probable error instead of the standard error, the formula would be:

$$\text{P.E.}_M = .6745 \frac{\sigma}{\sqrt{N}}$$

And the substitutions would be

$$\text{P.E.}_M = .6745$$

$$= \pm .310$$

It is obvious that the only thing done was to multiply the standard error by the constant, .6745. The probable error gives the range within which the chances are 1 to 1 that the mean of any other sample will not be less than 100.890 nor greater than 101.510. We have simply included a smaller proportion of the area of the frequency polygon within the limits of the measure of error. The standard error and the probable error are not to be contrasted. The first simply accounts for the range within which two-thirds of the cases will likely fall, whereas the other accounts for the range within which one-half of the cases will likely fall.

Similarly the standard error and the probable error of the median or of either quartile may be determined by multiplying the standard error by the appropriate constant:

$$\text{S.E.}_{Md} = 1.2533 \frac{\sigma}{\sqrt{N}} \qquad \text{P.E.}_{Md} = .8454 \frac{\sigma}{\sqrt{N}}$$

$$\text{S.E.}_{Q1} = 1.3626 \frac{\sigma}{\sqrt{N}} \qquad \text{P.E.}_{Q1} = .9191 \frac{\sigma}{\sqrt{N}}$$

The standard error and the probable error for the third quartile are the same as for the first quartile.

The standard error of the standard deviation for a distribution conforming to the normal curve is:

$$\text{S.E.}_{\sigma} = \frac{\sigma}{\sqrt{[illegible]}}$$

Substituting the appropriate values from the intelligence test data, we have:

$$\text{S.E.}_{\sigma} = \frac{18.8}{\sqrt{(2)\,(1671)}}$$
$$= .325$$

Thus, the standard deviation should be written 18.8 ± .325. The chances are 2 to 1 that any other sample selected would have a standard deviation between 18.475 and 19.125. This formula is accurate only for a normal distribution. For a skewed distribution the following formula may be used:

$$\text{S.E.}_{\sigma} = \mu_4 - \mu_2^2 \ [illegible]$$

$$= .0323, \text{ intervals}$$
$$= .323, \text{ points}$$

This standard error of the standard deviation differs somewhat from the other. That is to be expected, because we have previously shown that the present sample of I.Q's varies considerably from a normal distribution.

In Chapter XI the formulas for computing standard errors of regression curves were given and illustrated. Also the probable error of a coefficient of correlation was illustrated.

The standard error of a coefficient of correlation—simple, multiple, or partial—is determined by the following formula:

$$\text{S.E.}_{r} = \frac{1 - [illegible]^2}{[illegible]}$$

If the right side of this equation is multiplied by .6745, the result is the probable error of the coefficient of correlation. This formula is less accurate for distributions which depart widely from normal.

One other measure of standard error is important, and that is the measure of the significance of variability between two rates, such as per cent, per mille, per hundred thousand, etc. A recent paper by Professor Frank A. Ross[3] has emphasized the importance of calculating the standard error of rates in ecological studies of social phenomena. The formula in general use for computing this measure is as follows:

$$\frac{R(b - R)}{N}$$

in which

σ_R = the standard error of a rate
R = the rate—per cent, per mille, etc.
b = the base—100, 1000, etc.
N = population

Suppose crime rates have been computed for two census tracts in a city, one being near the central business district and the other at some distance from this locality. The rates may be based upon a relatively small number of cases, and they may differ considerably. If conditions remained the same in the two tracts, would we expect similar differences in rates to occur in another year? As Professor Ross points out, other questions arise here besides that of scarcity of data, but the probable variability of the difference between two rates, due to number of cases, may be determined from the following formula:

If

then the difference is significant and may be expected to recur under similar conditions. If

the difference is not significant, either because none really exists or

[3] Ross, Frank A., "Ecology and the Statistical Method," *Amer. Jour. Soc.*, Jan., 1933, pp. 508-517.

because the number of cases is too small to be reliable. The final formula is

If the result from the use of this formula is found to be less than the observed difference, then the observed difference is probably significant and may be expected to persist in the same direction under similar conditions in other years.

In applying the principle of standard error, or probable error, the student should not assume that this mechanical test rules out all other considerations of adequacy and reliability of the sample. These measures are applicable only for reasonably large numbers. Mills suggests that if the number of items falls below 15 the formulas for standard errors should not be applied; in the case of correlation, he raised the minimum to 25. Even then the results do not warrant great confidence.[4] The application of these formulas assumes that, if successive samples were taken at random, the statistical measures secured would be distributed according to the normal law of error, that is, the normal curve. This assumption holds when the number of items is large and the samples are random. Yule warns against an easy assumption that the sample is large enough to insure reliability. He says, ". . . if n is small, the rule that a range of three times the standard error includes the majority of the fluctuations of simple sampling *of either sign* does not strictly apply, and the 'probable error' becomes of doubtful significance."[5] The adequacy of the sample must always be determined by the investigator.

The errors referred to above are known as "errors of simple sampling," that is, errors due to chance when all precautions have been taken to obtain a random sample. But errors in sampling, aside from fluctuations due to simple sampling, cannot be accounted for by the formulas given. Fluctuations in the sample due to bias or inaccurate collection of data are not indicated by measures of standard error. These are matters of common sense and careful work.

5. EXERCISES

1. Toss 10 pennies 500 times, keep a record of the heads at each throw, and compare the results with the expansion of the binomial, $(p + q)^{10}$.

[4] See Mills, *op. cit.*, pp. 559, 560.
[5] *Op. cit.*, p. 353.

(a) Compare the standard deviation of the experimental data with the standard deviation of the theoretical distribution.
(b) Compare your distribution of heads with that obtained by Weldon (see Table LXXX).

The following table gives the hourly output of 14 women button workers in a factory over a period of 4 weeks to 4 months, showing the production per hour in intervals of .2 of a pound and the frequency of the occurrence of production at each class-interval:

TABLE LXXXVII

HOURLY PRODUCTION AND FREQUENCY OF PRODUCTION IN EACH INTERVAL—BUTTON WORKERS [1]

Class-Interval in Pounds	Frequency
Total	2,080
Below 1.4	23
1.4–1.6	35
1.6–1.8	59
1.8–2.0	128
2.0–2.2	245
2.2–2.4	319
2.4–2.6	351
2.6–2.8	322
2.8–3.0	252
3.0–3.2	194
3.2–3.4	101
3.4–3.6	35
Above 3.6	16

[1] Florence, P. S., *The Statistical Method in Economics and Political Science*, p. 70. New York: Harcourt, Brace & Co., 1929.

(a) Determine the mean rate of production and its standard deviation.

(b) Compute the first to fourth moments of this distribution and determine the values of β_1 and β_2. How do these functions compare with the corresponding functions of a normal distribution? Would you conclude that piece-work rates follow the normal law of error?

(c) Assuming 2,080 items and the standard deviation found, compute the values of the ordinates for a normal distribution at intervals of .2 of the standard deviation. Make a graph of the actual and theoretical distributions. Does the normal curve appear to fit closely?

(d) Assuming 2,080 items, redistribute them for a normal distribution, using the table of integrals computed in terms of area.

Make a graph of the actual and the theoretical distribution of the 2,080 items. Does the normal curve appear to fit closely?

(e) What is the standard error of sampling? Apply it to two or three different class frequencies.

(f) Apply the Chi-Square test for goodness of fit to the piece-work data. For the probability integral of your result consult Appendix A, Table CXXIII.

(g) Determine the standard errors of the mean and the standard deviation. What do these errors tell you about the sample?

3. Let each student find a group of data which seem to conform to the normal curve and compute all the statistical measures applied to the piece-work data. These may be secondary data published in some book, or they may be primary data gathered by the student. This exercise should give the student practice in estimating the form of a frequency distribution.

6. REFERENCES

Kelly, T. L., *Statistical Method*, Chap. V.
Mills, F. C., *Statistical Methods*, Chaps. XV, XVI.
Pearl, Raymond, *Medical Biometry and Statistics*, Chaps. X-XII.
Rietz, H. L., *Handbook of Mathematical Statistics*, Chap. V.
Weld, L. D., *Theory of Errors and Least Squares*, Chaps. II-IV.
Yule, U. G., *An Introduction to the Theory of Statistics* Chaps. XIII-XV.

CHAPTER XIII

Time Series

I. INTRODUCTION

THE most common characteristic of social data is that they vary in time. The chronological changes in the quantity and quality of social data are especially significant, because we want to know whether certain conditions are recurrent and what their general tendency of development is. Population facts, marriage, divorce, births, deaths, crime, insanity, poverty, and any number of other series of social data occur in time. Both private and governmental reports of social facts present them as having occurred in certain months or years. It is not enough to have the raw data classified and put into tables; they must be analyzed to extract the meaning that is most significant for an understanding of society and for determining social policy. Special methods of analyzing time series have been developed, and it is the object of this chapter to describe and illustrate the more usual methods.

Before turning to the technical procedures, however, attention may be directed to the logic of time as a category in social statistics. Social facts change in time, but man has developed ways of charting the passage of time. He has set guideposts along the route of human history, and he has worked out certain measuring sticks which enable him to know how much time has elapsed between one social event and another. Some of the measuring sticks are based upon astronomical observations. The earth's relation to the sun determines certain physical recurrences, which are the effects of the revolution of the earth about the sun and of the declination of the earth's axis. These physical conditions determine seasons, and a little thought will show how large is the number of social facts affected by the seasons. The rotation of the earth on its axis determines night and day, and many social phenomena vary as the result of this fact. In the process of adaptation to his

physical and biological environment man in many ways adapted his culture patterns to these astronomical recurrences. Religious observances have a definite relation to seasons of the year. Production and consumption habits are notoriously seasonal in their variation, or we should not have such widespread efforts to stabilize employment. Man devised tools for measuring the length of day and year. Weeks and months are different types of temporal units; they are purely matters of culture, and their lengths are only remotely related to astronomical observations. Of course, all the measuring tools for time are matters of culture, but some of them divide physical recurrences into definite pieces, such as seconds, minutes, hours, and years. Time as duration is not a cause of social variation, but physical and social facts undergo change because of the interaction of forces in apparently unstable equilibrium in nature and society. These forces act in time, and it is the resultants of their successive actions that the social statistician wants to record and analyze. Hence, he adopts the conventional units of time as a sort of jointed clothes-line upon which to hang social facts at regular intervals. This brings one kind of order into the mass of data, and then he can proceed to study the quantitative variations occurring at different points along the clothes-line. Things grow and endure for a certain piece of time, and then they wear out or disintegrate—even human beings; the social statistician wants to know how big a piece of time is required for forces to develop and wear out a human being, a dynasty, a nation, or a culture.

Observation has shown several different kinds of temporal variations. One of these is called secular, or long-time, trend. In social statistics secular trend is the general direction of growth or decline of a series of social phenomena over a period of 10, 25, 100 or more years. The trend may be in a straight line, or it may be curvilinear. The duration of a "secular trend" is relative to what in a human life seems to be a "long time." It is a practical concept. In fact, we do not have data sufficiently complete for any social series to describe its absolute secular trend. Even the logistic curve, describing the secular trend of population growth, involves a speculative analogy between the life of an organism and the life of a human race. But for practical purposes we can speak of the secular trend of per capita wealth in a nation, the production of automobiles, divorce, crime, or employment. Secular trend is measured in terms of the average amount of change per month or year over a long period of time. On this basis estimates may be made

of probable values in succeeding years, though such estimates, known as extrapolation, are not reliable if carried far in advance of the last actual data. Because the secular trend is not likely to show sharp variations within a short period of time, its computation is often an aid to social planning. For example, the secular trend of the number of children of high school age over a period of years would aid school administrators in planning building construction several years in advance.

Seasonal fluctuations are another type of temporal change, recurring in wavelike fashion each year. They may be caused by something in the physical environment, or they may be due to cultural habits or to seasonal fluctuations in some other social series. One of the social series best known for its seasonal fluctuations is employment. Certain industries, such as building construction, seem to be limited by climatic conditions to operating on full time during the warm months of the year and on part time during the cold months. The packing and canning industries have sharp fluctuations in the numbers employed because of the fluctuations in the flow of livestock and vegetables. But death rates also show marked seasonal variations. The attendance at theaters and churches has regular ups and downs during the year. Charitable relief goes up in the winter and down in the summer. It is important to measure the extent of such seasonal fluctuations so that plans may be made to meet them as effectively as possible. Efforts at the stabilization of employment are directed toward eliminating seasonal fluctuations in production, and in order to accomplish this it is necessary to understand the seasonal fluctuations of all of the factors determining seasonal changes in the industry concerned.

Besides secular trend and seasonal fluctuations, there are cyclical variations in social series. These occur at longer intervals than seasonal changes but are relatively short as compared with the secular trend. The most commonly recognized cyclical variations are those shown by business: the booms and the depressions. From the peak of one boom to the peak of another may be several years, and this period constitutes a cycle. Many social series, such as poverty and crime, are correlated with cyclical variations in business. If we think of secular trend as a straight line or a parabola, then the cyclical variations represent oscillations above and below the trend line. They also are wavelike, but the amplitude of the waves is greater than for seasonal fluctuations. Cyclical variations are extremely complex in their origin; they seem to result from an intricate interaction of a number of social or economic condi-

tions, over which no control has been achieved. Cyclical unemployment is one of the greatest of social problems, but as yet no way has been found to reduce its severity. More complete analysis of cyclical variations of different social and economic series may lead to such an understanding of the problems involved that control can be attained. Because of the seriousness to society of cyclical variations, it is particularly important for the student to know how to measure these changes in time series.

A fourth kind of temporal variation is known as residual variation. This is a term covering a multiplicity of irregular changes in social and economic phenomena. A change in some series may be due to an earthquake, to storms, to droughts, to a war, or to other forces operating at a particular time but not likely to recur at any predictable time. The residual changes are what remain after secular trend, seasonal fluctuations, and cyclical variations have been accounted for. In this volume, however, we are chiefly concerned with secular trend and seasonal and cyclical variations.

2. MEASUREMENT OF SECULAR TREND

Before proceeding to the computation of the secular trend of a series of data, the investigator should decide whether, in order to answer his question, allowance should be made for such factors as population change, change in the age ratios, or fluctuations in the general price level. The secular trend of actual dollars expended for the operation of the United States government would be quite different from the secular trend of actual expenditures adjusted for changes in the general price level. Likewise a phenomenon like divorce shows differences, when the gross number is used and when divorces are expressed as so many per 1,000 marriages. In 1889 there were 31,735 divorces in the United States, and in 1928 the number was 192,342.[1] That is an increase of 606.1 per cent. For the same years the divorce rates per 1,000 marriages were respectively 60 and 166, or an increase in the *rate* of only 276.6 per cent. The secular trend for actual divorces would be much more sharply upward than the secular trend for rates per 1,000 marriages. The investigator must decide which data are best suited to his purpose: divorces or divorce rates. If he is interested in the absolute increase in divorces, then he would want to know the secular trend of the number of divorces granted; if he is concerned with the relative increase in the rate of divorce, he would

[1] Reuter, E. B., and Runner, J. R., *The Family*, p. 211. New York: McGraw-Hill, 1931. Quoted from *Statistical Abstract of the United States*, 1929, p. 91.

want the secular trend of annual divorce rates. Whenever the secular trend of a social series is to be determined, the decision must be made as to whether or not interest is in relative or in absolute variations.

Secular trend may be computed in a number of ways. The first to be presented is a graphic method, and the data used are divorce rates for Indiana from 1899 to 1928:

TABLE LXXXVIII

DIVORCES PER 100,000 POPULATION IN INDIANA, 1899 TO 1928[1]

Year	Divorce Rate
1899	144
1900	143
1901	143
1902	147
1903	155
1904	134
1905	147
1906	154
1907	157
1908	160
1909	157
1910	172
1911	180
1912	201
1913	189
1914	181
1915	187
1916	198
1917	198
1918	194
1919	207
1920	221
1921	212
1922	238
1923	247
1924	239
1925	245
1926	246
1927	256
1928	248

[1] Data partly from *Marriage and Divorce, 1927*, and *Marriage and Divorce, 1929*, United States Bureau of the Census, and partly provided by Professor Charles R. Metzger, of Indiana University.

An examination of the table shows that the trend of the divorce rate is upward, but the statistical problem is to fit a trend line to the data. Is the trend linear or curvilinear? It appears to be linear. In order to get a picture of the distribution of divorce rates Figure LXV was drawn. The solid line connects the tops of the ordinates of the divorce rates:

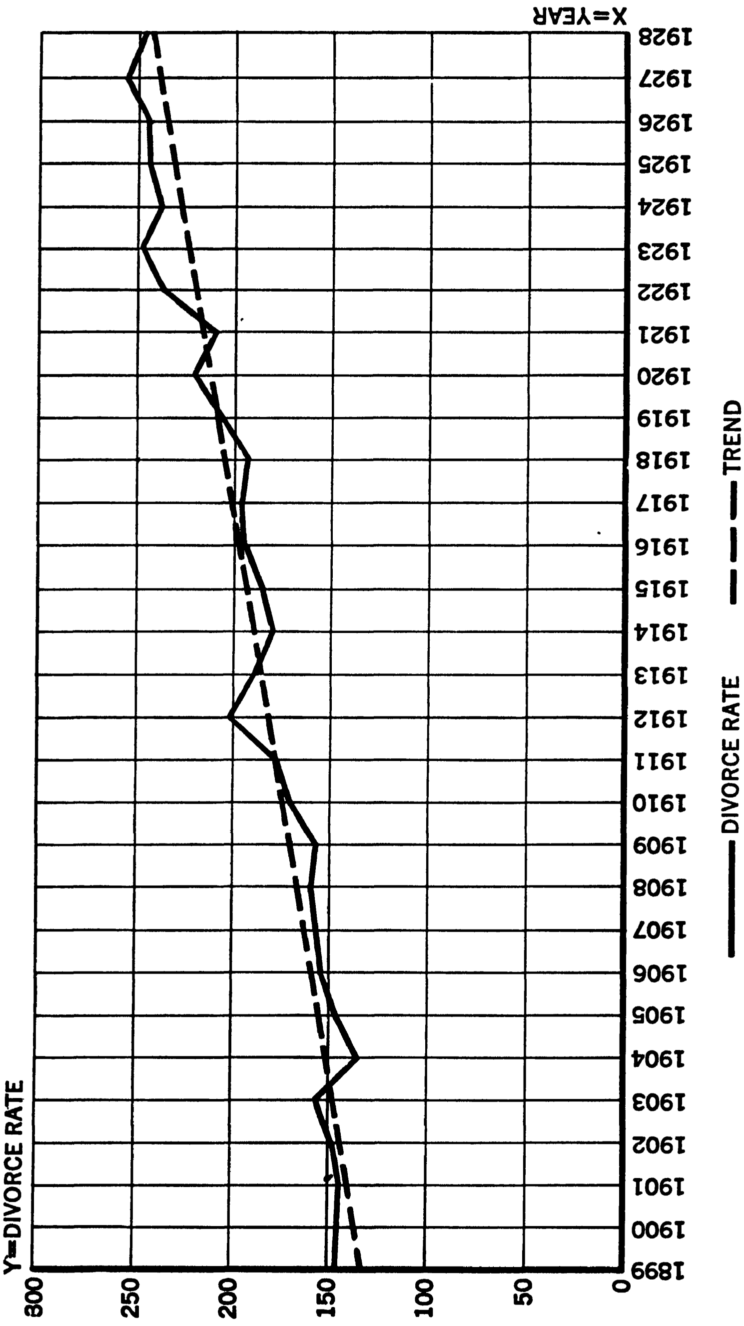


FIGURE LXV.—TREND OF DIVORCE RATES IN INDIANA, 1899-1928

The broken line is the line of trend fitted by the method of semi-averages. The mean divorce rate for the first 15 years was determined, and the circle on the ordinate for 1907 marks it. The mean rate for the second 15 years was found and is indicated by the circle on the ordinate for 1921. These two semi-averages were connected by a straight line, and the line was prolonged in each direction, to 1899 and to 1928. The straight line fits the data rather well; there is not much question of curvilinearity. The trend cuts the 1899 ordinate at 133 and the 1928 ordinate at 243. Subtracting 133 from 243, we get 110. If 110 is divided by 30, we get 3.7 as the average increase per year in the divorce rate; that is, the annual trend value is 3.7. If the trend line were projected to 1929, we would add 3.7 to 243 making 246.7. That

TABLE LXXXIX

MOVING AVERAGES OF DIVORCE RATES

Year (1)	Annual Rates Y (2)	Four-Year Moving Average Y' (3)	Four-Year Moving Average Centered Y' (4)	Five-Year Moving Average Y' (5)	Seven-Year Moving Average Y' (6)
1899	144				
1900	143				
		144			
1901	143		146	146	
		147			
1902	147		146	144	145
		145			
1903	155		146	145	146
		146			
1904	134		147	147	148
		148			
1905	147		148	149	151
		148			
1906	154		152	150	152
		155			
1907	157		156	155	154
		157			
1908	160		160	160	161
		162			
1909	157		165	165	166
		167			
1910	172		172	174	174
		178			
1911	180		182	180	177
		186			
1912	201		187	184	181
		188			
1913	189		189	188	187
		190			
1914	181		190	191	191
		189			
1915	187		190	191	193
		191			
1916	198		193	192	193
		194			
1917	198		197	197	198
		199			
1918	194		202	204	202
		205			
1919	207		207	206	210
		209			
1920	221		215	214	217
		220			
1921	212		225	225	223
		230			
1922	238		232	231	230
		234			
1923	247		238	236	235
		242			
1924	239		243	243	240
		244			
1925	245		246	247	246
		247			
1926	246		248	247	
		249			
1927	256				
1928	248				

would be the trend value for 1929 and would be an estimate of the probable divorce rate in that year. The method of semi-averages is easy to use and requires little arithmetical work, but it is less exact than other methods of fitting the trend line.

Another method is called the method of the moving average. This is illustrated in Table LXXXIX (see preceding page). Since, in using a moving average to measure the secular trend, one is not always sure how many years to use, it is necessary to try several intervals. The moving averages for four years, four years centered, five years, and seven years are shown. The first average for the four-year interval is based upon the first four rates and is written halfway between the rate for 1900 and that for 1901. Any moving average for an even number of years would fall between two years; any moving average for an odd number of years falls in the middle of some year. Therefore, in order to make the even-year moving averages comparable with the odd-year moving averages it is necessary to take a second step and "center" the four-year moving average. This is done by adding the first two four-year averages and dividing by two, which gives 145.5, but since the nearest whole number is used, the centered moving average is written 146. The four-year moving average is computed as follows:

$$\frac{144 + 143 + 143 + 147}{4} = \frac{577}{4} = 144$$

$$143 + 143 + 147 + 156 = 589 = 147$$

Or the second, third, etc., averages may be found by a short cut: add to each average, such as 144, the $\pm$ difference between the number dropped and the number added (divided by 4), thus,

$$\text{Second Average} = 144 + (156 - 144)/4 = 147$$

It will be noted that to get the second average the first rate is dropped, the other three are retained in the second sum, and a new one is added at the end. This is the process by which each moving average, of whatever interval, is determined. The four-year moving average is centered in the same way but by adding only two of the four-year averages.

The differences between the various moving averages can be seen better in a graph, and this will also reveal which average seems to fit the data best. To present the averages in graphic form

it will be necessary to drop certain years at the beginning and at the end of the period, because we cannot have a seven-year moving average nearer the beginning than 1902 nor nearer the end than 1925. Figure LXVI shows the three moving averages and the actual data.

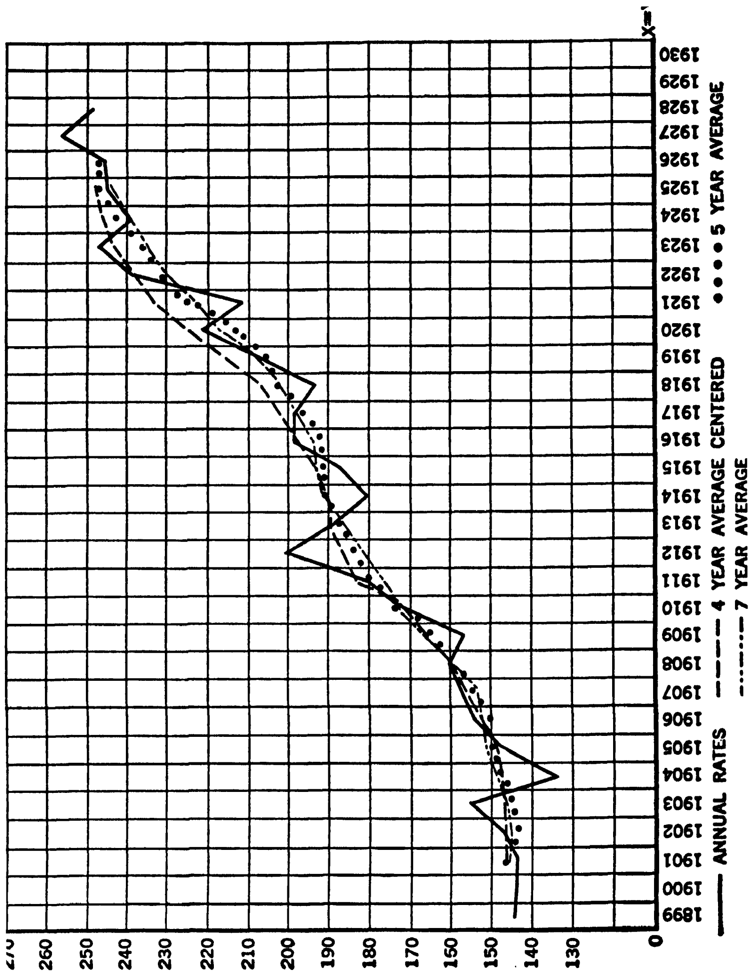
The three moving averages appear to fit about equally well, though the seven-year average is probably the best. Mills has shown that the best moving average for a series of data is one equal to the length of the cycle, to a multiple of the cycle, or to a period greater than the cycle. The cycles for the divorce rates vary somewhat, and that makes more difficult a decision as to the length of time required for the moving average. The cycle is usually, but not always, about five years in length. That moving average which reduces the number of cycles to a minimum is the best fit.[2] That is to say, the moving average which approaches nearest to a straight line and at the same time best fits the data is the one to use. For the divorce rates the four-year average shows 2 completed cycles and the beginning of a third. The five-year average shows about 2½ cycles. The seven-year average shows 2 cycles, while at the same time it fits the data very closely. Hence, we conclude that the seven-year moving average is the one to use in this case.

If the trend is curvilinear, a new difficulty arises. The trend of a series which is concave upward presents one problem, and the trend of a series which is concave downward presents an opposite problem. A moving average of a series with upward concavity will always exceed the actual trend values, whereas a moving average of a series with downward concavity will be smaller than the actual trend values. The moving average is not a good method of fitting a trend line to non-linear data, but, if it is to be used, "the period of the average should be the shortest which will serve to average out the cycles; equal, that is, to the average length of one cycle."[3] If the concavity is slight, the errors are naturally less than for series showing marked concavity. The flexibility of the moving average gives it an advantage as a measure of trend over certain other measures, though for some purposes it is not as useful as a mathematical curve.

Some data show a sufficiently definite and consistent trend to

[2] For a demonstration of the moving average of best fit, see Mills, *op. cit.*, pp. 260-265.

[3] *Op. cit.*, p. 267. See also pp. 265-267 for demonstration of error in moving averages for curvilinear series.



GURE CE FOU TE VE,

justify fitting to them a mathematical instead of a moving average curve. If the same forces operate over a long period of time to produce the changes occurring in the series, the trend is likely to be of this definite type. Where additional forces enter to affect variation during the period, the changes in the series are likely to be irregular and will be more adequately represented by a moving average than by a mathematical curve. For example, the State of South Carolina does not allow divorce on any grounds. If the law were amended to permit divorce on one ground, a record of divorce cases would appear in the state. If somewhat later several other grounds were permitted for divorce, the curve would doubtless show a sharp turn upward. It is such irregularities that make it inadvisable to fit mathematical curves to some social data, though, of course, there are series to which such a curve may be fitted with accuracy.

Trend as indicated by a moving average is obviously empirical; it assumes no law of growth. But a mathematical curve is a method of stating a law of change. As a matter of fact, mathematical curves fitted to social data are also empirical, but they imply greater certainty concerning trend than does a moving average. Because of this empirical character, the implications of a mathematical curve should be definitely hedged about with cautions. In the present state of the development of the social sciences we cannot state laws in the sense that they can be stated in the natural sciences. Too many factors are either unknown or cannot be taken into account because of their qualitative nature. Nevertheless, some of the methods of fitting mathematical curves to social data can be illustrated for purposes of experimentation on the part of the student. As more reliable data accumulate, approximations to laws of change may be discovered and stated with accuracy in mathematical terms.

In order to show the varying degrees of fit in different curves, we shall use the divorce data for illustrating the computation of mathematical curves. Three mathematical curves will be fitted to the divorce data: a straight line, a second degree parabola, and a logarithmic curve. Then the curves will be compared with each other and with the seven-year moving average. A straight line will be fitted to the data first by the method of least squares. The general equation for the line will be $Y = a + bX$. The problem is to compute the values of a and b, and the method is shown in Table XC.

TABLE XC

FITTING A STRAIGHT LINE TO THE DIVORCE DATA

Year	Number of the Year X	Divorce Rates Y	X^2	XY	Estimated Values—Trend Y'
1899	1	144	1	144	127.6
1900	2	143	4	286	131.9
1901	3	143	9	429	136.2
1902	4	147	16	588	140.5
1903	5	155	25	775	144.8
1904	6	134	36	804	149.1
1905	7	147	49	1029	153.4
1906	8	154	64	1232	157.7
1907	9	157	81	1413	162.0
1908	10	160	100	1600	166.3
1909	11	157	121	1727	170.6
1910	12	172	144	2064	174.9
1911	13	180	169	2340	179.2
1912	14	201	196	2814	183.5
1913	15	189	225	2835	187.8
1914	16	181	256	2896	192.1
1915	17	187	289	3179	196.4
1916	18	198	324	3564	200.7
1917	19	198	361	3762	205.0
1918	20	194	400	3880	209.3
1919	21	207	441	4347	213.6
1920	22	221	484	4862	217.9
1921	23	212	529	4876	222.2
1922	24	238	576	5712	226.5
1923	25	247	625	6175	230.8
1924	26	239	676	6214	235.1
1925	27	245	729	6615	239.4
1926	28	246	784	6888	243.7
1927	29	256	841	7424	248.0
1928	30	248	900	7440	252.3
Total	465	5700	9455	97914	

$$M_x = \frac{465}{30} = 15.5$$

$$M_y = \frac{5700}{30} = 190.0$$

$$b = \frac{\Sigma XY - nM_xM_y}{\Sigma X^2 - n(M_x)^2} = \frac{97914 - 30(15.5)(190.0)}{9455 - 30(15.5)^2} = 4.3$$

$$a = M_y - bM_x = 190.0 - 4.3(15.5) = 123.3$$

$$Y = 123.3 + 4.3X$$

The trend line is determined by assuming values of X successively from 1 to 30, that is, using the first year of the period, the second year, etc., as values of X. The trend values are given in the table.

As suggested when it was fitted by the method of semi-averages, the straight line gives a fairly close fit; the differences between the actual and the trend values are not great.

But it may be that a closer fit could be obtained by the use of a second degree parabola. Table XCI shows the method:

TABLE XCI

COMPUTATION OF PARABOLIC CURVE

Year X (1)	Rate Y (2)	X^2 or U (3)	XU or X^3 (4)	U^2 or X^4 (5)	XY (6)	UY or X^2Y (7)	Trend Values Y' (8)
1	144	1	1	1	144	144	127.5
2	143	4	8	16	286	572	132.0
3	143	9	27	81	429	1287	136.4
4	147	16	64	256	588	2352	140.8
5	155	25	125	625	775	3875	145.2
6	134	36	216	1296	804	4824	149.5
7	147	49	343	2401	1029	7203	153.9
8	154	64	512	4096	1232	9856	158.2
9	157	81	729	9561	1413	12717	162.6
10	160	100	1000	10000	1600	16000	166.9
11	157	121	1331	14641	1727	18997	171.2
12	172	144	1728	20736	2064	24768	175.5
13	180	169	2197	28561	2340	30420	179.8
14	201	196	2744	38416	2814	39396	184.1
15	189	225	3375	50625	2835	42525	188.3
16	181	256	4096	65536	2896	46236	192.8
17	187	289	4913	83521	3179	54043	196.9
18	198	324	5832	104976	3564	64142	201.1
19	198	361	6859	130321	3762	71478	205.3
20	194	400	8000	160000	3880	77600	209.5
21	207	441	9261	194481	4347	91287	214.7
22	221	484	10648	234256	4862	106964	217.3
23	212	529	12167	279841	4876	112148	222.1
24	238	576	13824	331776	5712	137088	226.2
25	247	625	15625	390625	6175	154375	230.3
26	239	676	17576	456976	6214	161564	234.4
27	245	729	19683	531441	6615	178605	238.6
28	246	784	21952	614656	6888	182864	242.7
29	256	841	24389	707281	7424	215296	246.8
30	248	900	27000	810000	7440	223200	250.9
465	5700	9455	216225	5276999	97914	2091826	

$M_x = 15.5$ $\qquad$ $M_y = 190.0$ $\qquad$ $M_u = 315.2$

The general form of the curve is $Y = a + bX + cX^2$, and the normal equations to be solved to determine the values of the constants are:

$$(\Sigma x^2)b + (\Sigma xu)c = \Sigma xy$$
$$(\Sigma xu)b + (\Sigma u^2)c = \Sigma uy$$

The terms in this equation are determined in the following manner:

$$
\begin{aligned}
\Sigma x^2 &= \Sigma X^2 - n(M_x)^2 = 9455 - 7207 = 2248 \\
\Sigma xu &= \Sigma XU - nM_xM_u = 216225 - 146568 = 69657 \\
\Sigma xy &= \Sigma XY - nM_xM_y = 97914 - 88350 = 9564 \\
\Sigma u^2 &= \Sigma U^2 - n(M_u)^2 = 5276999 - 2980531 = 2296468 \\
\Sigma uy &= \Sigma UY - nM_uM_y = 2091826 - 1796640 = 295186
\end{aligned}
$$

Substituting these values in the normal equations, we have:

$$
\begin{aligned}
&\text{(I)}\ 2248b + 69657c = 9564 \\
&\text{(II)}\ 69657b + 2296468c = 295186
\end{aligned}
$$

These equations must now be solved simultaneously. The Doolittle method will be used. Equation (I) will be divided through by the coefficient of b in the first equation with the sign changed, and then it will be set down with the derived equation (I′) below it:

$$
\begin{aligned}
&\text{(I)}\ 2248b + 69657c = 9564 \\
&\text{(I}'\text{)}\ -b - 30.99 = -4.25
\end{aligned}
$$

Equation (II) is then set down, and under it is written equation (I′) which has been multiplied by the coefficient of c in equation (I):

$$
\begin{array}{lrl}
 & \text{(II)} & 69657b + 2296468c = 295186 \\
\text{Adding,} & (69657)\,\text{(I}'\text{)} & \underline{-69657b - 2158670c = -296042} \\
 & & \qquad\qquad 137798c = -\ 856 \\
 & & \qquad\qquad\qquad c = -.006
\end{array}
$$

Substituting this value of c in either equation (I) or equation (II), we find the value of b:

$$b = 4.44$$

With these values known we can now determine the value of a by substituting the appropriate values in the following equation:

$$
\begin{aligned}
a &= M - bM_y - cM_u \\
&= 190.0 - 4.44(15.5) - (-.006)\ (315.2) \\
&= 123.1
\end{aligned}
$$

The equation of the curve can now be stated:

$$Y = 123.1 + 4.44X - .006X^2$$

The trend values in column (8) of Table XCI were determined by successively substituting values of X from 1 to 30. Since these values do not vary widely from the original data, it is possible to use this line of trend. But before a comparison is made between

the various lines of trend computed, we shall fit one more curve to the data, a logarithmic curve:

$$\text{Log } Y = a + b \log X$$

Table XCII shows the method:

TABLE XCII

COMPUTATION OF LOGARITHMIC CURVE

Year X (1)	Rate Y (2)	Logarithm of X $\overline{X}$ (3)	Logarithm of Y $\overline{Y}$ (4)	$\overline{X}\overline{Y}$ (5)	$\overline{X}^2$ (6)	Trend Values Y' (7)
1	144	.0000	2.1584	.0000	.0000	111.3
2	143	.3010	2.1553	.6487	.0906	128.5
3	143	.4771	2.1553	1.0283	.2276	139.7
4	147	.6021	2.1673	1.3049	.3625	148.3
5	155	.6990	2.1903	1.5310	.4686	155.3
6	134	.7782	2.1271	1.6553	.6056	161.3
7	147	.8451	2.1673	1.8316	.7142	166.5
8	154	.9031	2.1875	1.9755	.8156	171.2
9	157	.9542	2.1959	2.0953	.9105	175.4
10	160	1.0000	2.2041	2.2041	1.0000	179.2
11	157	1.0414	2.1959	2.2868	1.0845	182.8
12	172	1.0792	2.2355	2.4126	1.1647	186.1
13	180	1.1139	2.2553	2.5122	1.2408	189.3
14	201	1.1461	2.3032	2.6397	1.3135	192.1
15	189	1.1761	2.2765	2.6774	1.3832	194.9
16	181	1.2041	2.2577	2.7185	1.4499	197.5
17	187	1.2304	2.2718	2.7952	1.5139	200.0
18	198	1.2553	2.2967	2.8830	1.5758	202.4
19	198	1.2788	2.2967	2.9370	1.6353	204.7
20	194	1.3010	2.2878	2.9764	1.6926	206.8
21	207	1.3222	2.3160	3.0622	1.7482	208.9
22	221	1.3424	2.3444	3.1471	1.8020	211.0
23	212	1.3617	2.3263	3.1677	1.8542	212.9
24	238	1.3802	2.3766	3.2802	1.9050	214.8
25	247	1.3979	2.3927	3.3448	1.9541	216.6
26	239	1.4150	2.3784	3.3654	2.0022	218.4
27	245	1.4314	2.3892	3.4199	2.0489	220.1
28	246	1.4472	2.3909	3.4601	2.0944	221.8
29	256	1.4624	2.4082	3.5218	2.1386	223.4
30	248	1.4771	2.3945	3.5369	2.1818	225.0
		32.4236	68.1028	74.4196	38.9788	

$$M_{\bar{x}} = 1.0808 \qquad M_{\bar{y}} = 2.2701$$

The values of a and b in the general formula may now be determined in the following manner:

$$\frac{\Delta_{\bar{y}}}{)^2} = \frac{74.4196 - 73.6057}{38.9788 - 35.0439} =$$

$$a = M_{\bar{y}} - bM_{\bar{x}} = 2.2701 - .2235 = 2.0466$$

The equation for the line of trend will then be:

$$\text{Log } Y = 2.0466 + .2068 \log X$$

Substituting successively the values of the logarithms of X, we determine the logarithms of Y, that is, the logarithms of the trend values. These values may then be looked up in a table of logarithms and the trend values in natural numbers determined. That has been done in column (7) of Table XCII. These trend values are obviously not a good fit. In the middle of the period they are considerably higher than the original data, and at each end they are much smaller.

We are now ready to compare the differences in the trend values

TABLE XCIII

COMPARISON OF TREND VALUES DERIVED BY A 7-YEAR MOVING AVERAGE, A STRAIGHT LINE, A SECOND DEGREE PARABOLA, AND A LOGARITHMIC CURVE

Year	Divorce Rate Y	7-Year Moving Average		$Y - a + bX$		$Y = a + bX + CX^2$		Log $Y = a + b \log X$	
		Y'	$Y-Y'$	Y'	$Y-Y'$	Y'	$Y-Y'$	Y'	$Y-Y'$
1899	144			127.6	16.4	127.5	16.5	111.3	32.7
1900	143			131.9	11.1	132.0	11.0	128.5	14.5
1901	143			136.2	6.8	136.4	6.6	139.7	3.3
1902	147	145	2	140.5	6.5	140.8	6.2	148.3	− 1.3
1903	155	146	9	144.8	10.2	145.2	9.8	155.3	− .3
1904	134	148	−14	149.1	−15.1	149.5	−15.5	161.3	−27.3
1905	147	151	− 4	153.4	− 6.4	153.9	− 6.9	166.5	−19.5
1906	154	152	2	157.7	− 3.7	158.2	− 4.2	171.2	−17.2
1907	157	154	3	162.0	− 5.0	162.6	− 5.6	175.4	−18.4
1908	160	161	− 1	166.3	− 6.3	166.9	− 6.9	179.2	−19.2
1909	157	166	− 9	170.6	−13.6	171.2	−14.2	182.8	−25.8
1910	172	174	− 2	174.9	− 2.9	175.5	− 3.5	186.1	−14.1
1911	180	177	3	179.2	.8	179.8	− .2	189.3	− 9.3
1912	201	181	20	183.5	17.5	184.1	16.9	192.1	8.9
1913	189	187	2	187.8	1.2	188.3	.7	194.9	− 5.9
1914	181	191	−10	192.1	−11.1	192.6	−11.6	197.5	−16.5
1915	187	193	− 6	196.4	− 9.4	196.9	− 9.9	200.0	−13.0
1916	198	193	5	200.7	− 2.7	201.1	− 3.1	202.4	− 4.4
1917	198	198	0	205.0	− 7.0	205.3	− 7.3	204.7	− 6.7
1918	194	202	− 8	209.3	−15.3	209.5	−15.5	206.8	−12.8
1919	207	210	− 3	213.6	− 6.6	214.7	− 7.7	208 9	− 1.9
1920	221	217	4	217.9	2.1	217.3	3.7	211.0	10.0
1921	212	223	−11	222.2	−10.2	222.1	−10.1	212.9	− .9
1922	238	230	8	226.5	11.5	226.2	11.8	214.8	23.2
1923	247	235	12	230.8	16.2	230.3	16.7	216.6	30.6
1924	239	240	− 1	235.1	3.9	234.4	4.6	218.4	20.4
1925	245	246	− 1	239.4	5.6	238.6	6.4	220.1	24.9
1926	246			243.7	2.3	242.7	3.3	221.8	24.2
1927	256			248.0	8.0	246.8	9.2	223.4	32.6
1928	248			252.3	− 4.3	250.9	− 2.9	225.0	23.0

found by the seven-year moving average, the straight line, the parabola, and the logarithmic curve. Table XCIII gives results. The mean deviations from the actual data, disregarding algebraic signs, are as follows:

7-Year Moving Average	5.8
Straight Line	8.0
Parabola	8.3
Logarithmic Curve	15.4

The mean of the deviations from the moving average is the smallest. It should be noted, however, that the mean of the deviations from the moving average is based upon 24 instead of 30 years, that is, 1902 to 1925. We may say, then, that the moving average gives the closest and the logarithmic curve the worst fit. The moving average is flexible, and this gives it a general advantage over other methods of smoothing time series. Its chief limitation lies in the fact that the larger the number of years included in the moving average period, the more years at each end of the series will be left without any average. Hence, the choice of a moving average or some other measure of trend will depend, not only upon the closeness of fit, but upon whether it is important to the problem to have an average for every year in the period.

3. MEASUREMENT OF SEASONAL FLUCTUATIONS

Seasonal fluctuations in social phenomena have been recognized by everyone. Besides the theoretical interest in understanding the amount of seasonal fluctuation in various types of social phenomena, there are important practical interests. In social planning it is necessary to know when seasonal fluctuations come and how great they are so that they may be taken into consideration. For example, the marked seasonal changes in the demands made upon charitable relief agencies have to be considered in apportioning the budget of such agencies in order properly to spread expenditures throughout the year. Mortality and morbidity also vary with seasons, and both private physicians and public health officers need to know what the seasonal fluctuations are for different diseases and for all diseases taken together. Students of climate in relation to human behavior have noted changes in the efficiency of workers under varying temperature and humidity, both of which have mean seasonal fluctuations. Another practical interest in this sub-

ject is the desire to eliminate the seasonal influence on social phenomena, when the principal interest is in the cyclical variations. Cycles cover longer intervals of time than seasonal recurrences, and, if the swing of these longer social changes is to be measured accurately, due allowance must be made for the regularly recurring seasonal variations. Otherwise one might interpret an upswing or a downswing of the curve as a cyclical variation, when in fact it was only the normal seasonal fluctuation and the direction of the cyclical change might be in the opposite direction. This is well illustrated by the level of employment. The cycle of employment may be going up in the winter months, but it is almost certain that the seasonal fluctuation is downward during that period in any year. If we are to make allowance for these different types of variation, we must have some way of measuring the quantity of seasonal change.

Several such methods have been proposed and used. For purposes of illustration mortality data for the State of Indiana from 1911 to 1930 will be used. Mortality rates per 1,000 population in the state are published each month. In Table XCIV these mortality rates have been changed a little in order to enlarge the figures dealt with; this makes variations more obvious. The rates per 1000 population were reduced to index numbers by expressing each monthly rate in terms of a percentage of the mean mortality rate in 1911; that is, 1911 was used as a base year.

Seasonal indexes for these data might be computed in either of the following ways: (1) by taking the mean rate of all January rates, of all February rates, etc., after which we would have percentage figures for each month, and the total for the 12 months would be 1,200; (2) by arranging the rates for each month in an array, or a multiple frequency table, and taking the mean of the 2, 4, 6, or more middle rates; (3) by Persons' chain-link-median method; (4) by Falkner's method of computing the ratio of the original data to the trend values and taking the adjusted median monthly values; (5) by the method of a twelve-month moving average, centered, in connection with the method of median monthly values. Methods (1), (2), and (4) will be illustrated. Method (3) is reliable and has been used extensively by the Harvard Committee on Economic Research, but it does not seem to have any advantage over methods (2) and (4), and it requires much more arithmetical work. Method (5) is easy to understand and to compute, but the same practical objection may be raised

TABLE XCIV

MORTALITY RATES IN INDIANA, 1911–1930, EXPRESSED AS PERCENTAGES OF THE MEAN MONTHLY RATE IN 1911

	Jan.	Feb.	Mar.	Apr.	May	June	July	Aug.	Sept.	Oct.	Nov.	Dec.
1911	114.6	113.0	113.8	111.6	94.5	85.6	102.4	92.9	87.3	93.7	91.2	97.7
1912	111.4	111.4	116.8	112.1	91.1	84.7	97.5	100.0	100.8	99.2	94.3	108.1
1913	111.2	113.7	121.8	108.0	98.4	101.5	100.8	108.0	99.2	96.8	96.8	91.2
1914	104.0	111.4	121.8	116.2	101.4	91.1	94.3	93.5	95.1	95.1	91.9	96.0
1915	104.0	122.5	125.0	107.1	89.5	81.5	85.5	84.7	92.7	88.7	93.5	104.8
1916	135.4	117.0	100.0	111.2	96.8	85.5	103.2	100.0	100.8	97.6	94.3	102.3
1917	120.1	137.0	132.2	116.0	111.4	94.3	96.0	98.3	103.2	98.3	95.2	105.6
1918	116.1	108.8	124.2	131.3	106.3	82.3	87.8	94.3	87.1	195.8	159.7	178.2
1919	133.8	112.0	146.7	103.1	87.8	75.0	87.1	78.2	77.3	82.3	84.7	96.0
1920	120.9	181.3	129.8	102.3	103.2	85.5	83.1	79.1	82.3	87.2	94.4	105.6
1921	105.6	96.0	98.4	92.0	94.3	82.3	94.3	92.7	88.0	96.0	91.2	96.0
1922	111.2	110.3	119.3	95.1	91.2	78.2	81.5	79.0	80.7	85.5	80.7	102.3
1923	117.0	131.4	130.6	110.4	94.4	88.0	81.5	89.6	89.6	90.4	83.8	98.5
1924	108.9	100.0	116.2	107.2	102.2	85.5	84.7	91.2	89.5	89.5	92.7	100.8
1925	98.4	101.6	121.0	109.7	88.0	87.2	91.2	90.4	83.1	97.6	93.5	103.2
1926	101.2	100.8	133.9	120.2	102.3	90.4	88.8	90.4	92.7	96.0	93.6	101.6
1927	104.0	94.4	100.8	101.6	94.4	93.6	92.7	87.2	88.8	95.2	94.4	100.8
1928	102.3	82.4	112.1	108.0	104.8	87.1	81.5	88.0	88.8	92.8	89.5	139.5
1929	143.6	103.2	108.8	97.6	98.4	87.2	84.7	84.7	87.9	88.8	88.8	105.6
1930	103.2	87.2	105.7	100.8	92.7	92.0	92.7	88.0	86.3	88.0	86.3	96.8

against it as against Persons' method. The other three methods illustrate simple ways of computing seasonal indexes, and (2) and (4) are highly reliable.

A simple method of getting a picture of seasonal fluctuations in a series of data is that of the multiple frequency table. The mortality rates are distributed in this manner in Table XCV. The

TABLE XCV

MULTIPLE FREQUENCY TABLE OF MORTALITY RATES SHOWING SEASONAL VARIATIONS

Rate	Month											
	J	F	M	A	M	J	J	A	S	O	N	D
Over 150		I								I	I	I
145 - 150			I									
140 - 145	I											
135 - 140	I	I										I
130 - 135	I	I	III	I								
125 - 130		I	I									
120 - 125	II	I	IIII	II								
115 - 120	II	I	III	III								
110 - 115	IIII	𝍸I	II	IIII	I							
105 - 110	II	I	II	IIII	I							IIII
100 - 105	𝍸I	IIII	III	IIII	𝍸	I	III	III	III			𝍸II
95 - 100	I	I	I	II	II		II	I	II	𝍸IIII	II	𝍸II
90 - 95		I		I	𝍸II	𝍸	𝍸	𝍸II	II	III	I 𝍸𝍸	
85 - 90		I			IIII	𝍸III	IIII	IIII	𝍸IIII	𝍸I	III	
80 - 85		I				IIII	𝍸II	II	III	I	III	
75 - 80						II		III	I			

seasonality of mortality rates appears to be definite: The rates are high in the winter months and low in the summer months; other seasons of the year show rates lying between these extremes, except that March seems to have the highest rates of any month.

This table may also be used in computing seasonal indexes by method (2).

Method (1) will be illustrated first. The means of the rates for the 20 Januaries, 20 Februaries, etc., are given in Table XCVI, along with the adjusted indexes of seasonal variation:

TABLE XCVI

COMPUTATION OF SEASONAL INDEXES FOR THE MORTALITY DATA BY METHOD (1)

Month (1)	Sum of Each 20-Mo. Group of Rates (2)	Monthly Average of Rates (3)	Monthly Averages Adjusted—Seasonal Index (4)	Monthly Variations from 100 Column (3) (5)
January	2266.9	112.3	112.4	12.4
February	2235.4	111.2	111.2	11.1
March	2378.9	118.3	118.3	18.3
April	2161.5	107.5	107.5	7.5
May	1943.1	96.7	96.8	− 3.2
June	1738.5	86.5	86.6	−13.4
July	1811.3	90.1	90.1	− 9.9
August	1810.2	90.1	90.1	− 9.9
September	1801.2	89.6	89.6	−10.4
October	1954.5	97.2	97.2	− 2.8
November	1890.5	94.0	94.0	− 6.0
December	2130.6	106.0	106.2	6.2
Total		1199.5	1200.0	
Mean		99.9	100.0	

The sum of the monthly mean rates is 1,199.5. A seasonal index is more convenient to use, if the sum is even 1,200.0. If each of the mean monthly rates is divided by the mean of all monthly rates, that is, 99.9, the adjusted averages are those given in column (4), and the total is 1,200.0, and the mean of the monthly indexes is 100.0. The relative values of the monthly indexes have not been changed by the adjustment. In column (5) the seasonal variations from the monthly average of 100.0 are given. It is quite clear that marked seasonal variations in death rates do occur. They range from 13.4 below to 18.3 above the monthly average.

Although method (1) is the simplest method for computing a seasonal index, it has one important weakness. It is characteristic of the mean to be affected unduly by the extreme variations, and the simple mean has been used to derive this index. Consequently, we should expect this index to exaggerate the seasonal variations. It may be used as a rough measure, but it is not as precise as a seasonal index may be made.

One other correction needs to be applied to the seasonal index, and that is the correction for secular trend. For example, mortality rates in Indiana have been declining during this 20-year period. The mean monthly decline should be added to the seasonal index to make this adjustment, because the secular trend is downward; if it were upward, the secular trend would be subtracted. Table XCVII gives the corrected monthly averages. The secular trend of the mortality rates is nonlinear; a second degree parabola fitted to the data gives a fairly close fit, of which the equation is:

$$Y = 102.5 + .292X - .0355X^2$$

The standard error of estimate is 5.99. When the annual trend values are computed by this formula, the mean annual decrease in the mortality rate is found to be .435, and the mean monthly decrease is .036. In order to correct for this amount of trend, using January as a base, .04 must be added to the February mortality rate, .08 to the March rate, .12 to the April rate, etc. When these additions have been made, the 12 monthly rates are added and an adjustment is made so that the indexes of seasonal variation equal 1,200.

TABLE XCVII

MONTHLY AVERAGES OF MORTALITY INDEXES CORRECTED FOR SECULAR TREND

Month (1)	Monthly Average of Mortality Indexes (2)	Monthly Averages Corrected for Trend (3)	Corrected Monthly Averages Adjusted to Equal 1200—Seasonal Index (4)	Monthly Variations from 100, Column (3) (5)
January	112.3	112.3	112.0	12.0
February	111.2	111.7	111.4	11.4
March	118.3	119.0	118.7	18.7
April	107.5	107.6	107.3	7.3
May	96.8	97.0	96.8	− 3.2
June	86.6	86.0	85.8	−14.2
July	90.1	90.3	90.1	− 9.9
August	90.1	90.4	90.2	− 9.8
September	89.6	89.9	89.7	−10.3
October	97.2	97.6	97.4	− 2.6
November	94.0	94.4	94.2	− 5.8
December	106.2	106.6	106.4	4.6
Total		1202.8	1200.0	
Mean		100.2	100.0	

There is an error in the final seasonal indexes as given in column (4) of Table XCVII due to the fact that the trend is non-linear.

The correction for trend is the average monthly decrease in mortality rates, but this involves an assumption of regular monthly decrements which would necessitate linearity of trend. Actually there is a slight acceleration in the rate of decrease. Hence, the correction to be added to December, when January is used as a base, is not exactly 11 times the correction added to February, but a little more than that because of the parabolic nature of the trend. In a similar manner the corrections for the other months are slightly erroneous.

These deviations from the monthly average for the mean year are of great value in estimating the actual relative importance of a mortality rate in any particular month. They show that every year there are normally months with rates higher than average and other months with rates normally lower than average. These variations reflect neither irregular causes of death nor the secular trend. They do show the regularly recurring variations during any year. Allowing for the possibility of unusual deviations, they indicate what is normally to be expected each month.

Method (2) eliminates the error due to the use of the mean monthly rates. It might be called the mean-median method, because the 20 Januaries, 20 Februaries, etc., are arranged in an array or a multiple frequency table, as Table XCV, and 2 or more of the middle values are added and the mean of these values is obtained. Hence, the extreme values exercise less influence on the final seasonal indexes than they do when method (1) is used. The mean-median method has been used by various writers in the computation of more than one kind of seasonal index, but Professor Chapin's use of it in connection with dependency indexes is especially pertinent to the interests of the social statistician. He has used it, along with measures of trend and cyclical variations, for the purpose of eliminating the seasonal factor in order to arrive at a measure of the residual variations in Minneapolis relief statistics.[4] Professor Chapin eliminates the seasonal variations from his dependency data by subtracting the seasonal factor from the original data. After which he removes the trend values, thus leaving only the cyclical and residual variations. For practical reasons he found this order of segregation desirable in his problem. Other writers have removed trend values from the data before

[4] See Chapin, F. S., "Dependency Indexes," *Social Forces*, Vol. V., No. 2, pp. 215-224.

TABLE XCVIII

THE MIDDLE FOUR MORTALITY RATES FOR EACH MONTH OF THE YEAR AND THEIR MEAN

Position of Rate in Array	Rates												
	Jan.	Feb.	Mar.	Apr.	May	June	July	Aug.	Sept.	Oct.	Nov.	Dec.	Totals
9th	105.6	108.8	116.2	107.2	94.4	85.5	87.1	88.0	88.0	92.8	91.9	100.8	1166.3
10th	108.9	110.3	116.8	108.0	94.4	85.6	87.8	89.6	88.8	93.7	92.7	101.6	1178.2
11th	111.2	111.4	119.3	109.7	94.5	87.1	88.8	90.4	88.8	95.1	93.5	102.3	1192.1
12th	111.2	111.4	121.0	110.4	96.8	87.2	91.2	90.4	89.5	95.2	93.6	102.3	1200.2
Total	436.9	441.9	473.3	435.3	380.1	345.4	354.9	358.4	355.1	376.8	371.7	407.0	4736.8
Means	109.2	110.5	118.3	108.8	95.0	86.4	88.7	89.6	88.8	94.2	92.9	101.8	

subtracting the seasonal factor. The order of elimination will depend upon the purposes of the investigator, but the net results should be substantially the same in both cases.

Since the mortality rates are for an even number of years, the median rate comes between the tenth and the eleventh year. If we had an odd number of years, the median rate would fall in the middle year of the array. In our illustration the seasonal index is determined by the mean-median method. In order to give equal weight to the rates on each side of the monthly medians, rates for the same number of years on each side of the median should be used. One, two, three, or more rates above and below the median might be taken. For the purpose of illustration two rates on each side of the median are used for obtaining the mean. Table XCVIII gives the four middle rates and the mean-median value for each month. It should be apparent from the construction of the multiple frequency table that the rates in each monthly column, or array, do not follow the same order of years. The rates for January at the middle of the array might be for entirely different years from those at the middle of the February array. But that does not affect the reliability of the results. The aim is to get a representative value for mortality rates for each month of the year. See Table XCVIII. (See Table XCV for complete data.)

TABLE XCIX

MEAN-MEDIAN RATES CORRECTED FOR TREND, ADJUSTED SEASONAL INDEXES, AND VARIATIONS FROM MONTHLY AVERAGE OF 100

Month (1)	Mean-Median Rates Corrected for Trend (2)	Seasonal Indexes (3)	Variations of Indexes from 100 (4) (4)
January	109.2	111.0	11.0
February	110.5	112.2	12.2
March	118.2	120.1	20.1
April	108.6	110.4	10.4
May	94.8	96.3	− 3.7
June	86.2	87.6	−12.4
July	88.4	89.8	−10.2
August	89.2	90.6	− 9.4
September	88.4	89.8	−10.2
October	93.8	95.3	− 4.7
November	93.5	94.0	− 6.0
December	101.3	102.9	2.9
Total	1181.0	1200.0	
Mean	98.3	100.0	

It remains to correct the means of the columns for secular trend and express these mean rates as seasonal indexes adjusted to equal 1,200 for the 12 months. Table XCIV shows these computations.

The computation of seasonal indexes by the method of the ratio of the actual rates to the trend rates is quite long compared with either of the preceding methods. In the first place, it involves computation of the monthly trend values. The annual trend values were computed by the parabolic equation given above (see p. 364). In view of the fact that the curvature is slight for any given year, we may for convenience assume that the trend line is straight and that a constant rate of decrease in death rates exists during the year. For example, the annual trend values, computed from the equation, will be centered at the middle of each year, because they are based upon 12-month averages. This should be shifted back to the middle of the first month of the year: January. The difference between this figure for January, 1911, and January, 1912, is found by subtracting 102.7 (January, 1912) from 102.9 (January, 1911). The decrease for the year is .2. If we carry it to one decimal place only, the rate for the first 6 months of the year will be assumed to be 102.9 for each month, and for the second six months it will be 102.8 for each month. For the first 6 months in 1912 the rate will be 102.7. Later years show more rapid declines in the rate, and at the end of the 20-year period it is declining at the rate of .1 each two months. After these computations are made, the ratio of each monthly actual value to each monthly trend value is computed and expressed as a percentage. When this has been done, the 20 Januaries, 20 Februaries, etc., are arranged in an array, and the mean of the middle four items is taken. This mean, when adjusted so that the 12 monthly means equal 1,200, is the seasonal index. This last step is clearly a mean-median method, but it has been applied after the trend has been removed by a more accurate method than was used in the other two illustrations. Table C gives the monthly mean-medians and the adjusted indexes.[5]

It will now be of interest to put the three types of seasonal indexes into one table, where comparisons can be made. In view of the fact that that the mean-median and the ratio-to-ordinate methods

[5] This method of computing seasonal indexes is known as the "ratio-to-ordinate" method and was developed by Dr. Helen D. Falkner. See "The Measurement of Seasonal Variation," *Journal of the American Statistical Association*, Vol. 19, pp. 167-179.

TABLE C

SEASONAL INDEXES COMPUTED BY THE RATIO-TO-ORDINATE METHOD

Month	Mean-Medians	Adjusted Indexes
January	109.6	110.6
February	108.7	109.6
March	117.1	118.1
April	108.0	108.9
May	96.4	97.2
June	87.4	88.2
July	91.1	91.9
August	91.7	92.5
September	90.9	91.7
October	94.1	94.9
November	92.6	92.4
December	103.1	104.0
Total	1190.7	1200.0
Mean	99.2	100.0

are the more accurate, the differences between these are indicated in the table. Table CI gives the three seasonal indexes:

TABLE CI

THREE SEASONAL INDEXES COMPARED—CORRECTED FOR SECULAR TREND

Month (1)	Method of Simple Means (2)	Method of Mean-Medians (3)	Method of Ratio-to-Ordinate (4)	(3) − (4) (5)
January	112.0	111.0	110.6	.4
February	111.4	112.2	109.6	2.6
March	118.7	120.1	118.1	2.0
April	107.3	110.4	108.9	1.5
May	96.8	96.3	97.2	− .9
June	85.8	87.7	88.2	− .5
July	90.1	89.8	91.9	−1.1
August	90.2	90.6	92.5	−1.9
September	89.7	89.8	91.7	−1.9
October	97.4	95.3	94.9	.4
November	94.2	94.0	92.4	1.6
December	106.4	102.9	104.0	−1.1

Disregarding signs, the mean difference between the mean-median and ratio-to-ordinate indexes is 1.3, whereas the mean difference between the simple mean and the ratio-to-ordinate method is 1.8, and the mean difference between the simple mean and the mean-median method is 1.3. These mean differences are fairly close. The correction for trend in the case of the mean-median method was the mean monthly decrease in the mortality index, which is

shorter than the correction for trend by the ratio-to-ordinate method. In view of these facts the mean-median method should be used as a time-saver, unless there are special reasons for preferring the ratio-to-ordinate method.

4. MEASUREMENT OF CYCLICAL FLUCTUATIONS

We are familiar with cyclical fluctuations chiefly through the discussion of business cycles which has been going on for some fifteen years. When the "business cycle" is mentioned, one immediately thinks of prosperity and depression in business. While more study has been given to cyclical variations by economists than by other social scientists, some work has been done on other social series. The interest in these latter cyclical variations appears to have developed out of the theory that economic conditions are correlated with a number of other social factors. The first work of this sort done in the United States was by Professor William F. Ogburn and Dr. Dorothy S. Thomas,[6] by G. P. Davies,[7] and by Miss K. E. Howland.[8] The work by Ogburn and Thomas is by far the most comprehensive. It considers the relation of the business cycle to marriages, divorces, births, deaths, and crime and tests the degree of relationship by means of correlation technique. Later Dr. Thomas pursued the study further in both the United States and England. In both countries she computed the degrees of correlation between the business cycle and marriages, births, deaths, pauperism, alcoholism, crime, and emigration. Changes in some of the social series lag behind changes in the business cycle, and allowance had to be made for this fact.[9] The aim in all of these studies was to measure the cyclical variations of social factors and then to compute the correlation between each series and the business cycle as the independent variable.

The short-time fluctuations called cycles may be computed for any social series varying in time. The term "cycle" implies recurrence. It suggests that variations go up for a while and then go down, and that these ups and downs recur with a fair degree of regularity. They are variations about the line of trend and are

[6] See their article, "The Influence of the Business Cycle on Certain Social Conditions," *Journal of the American Statistical Association,* September, 1922.

[7] "Social Aspects of the Business Cycle," *Quarterly Journal of the University of North Dakota,* January, 1922.

[8] "A Statistical Study of Poor Relief in Massachusetts," *Journal of the American Statistical Association,* December, 1922.

[9] Thomas, Dorothy S., *Social Aspects of the Business Cycle.* London: Routledge, 1925.

measured from that line, whether it be linear or curvilinear. Before determining the cyclical fluctuations, the seasonal factor should be removed from the data. If the data are annual, instead of monthly, the seasonal fluctuations do not appear at all. Under such circumstances, it is easy to compute a line of trend and then subtract the trend values from the actual values. The remaining variations will not be explained entirely by cyclical variations, because special causes intervene to produce residual variations. Professor Chapin has shown how these residual factors may be determined for dependency data. He removed the seasonal, trend, and cyclical factors and then found that some fluctuations still remained. These were the residuals and represented the effects of a multiplicity of minor causes. The residuals, he found, were distributed approximately in the form of a normal curve.[10] However, the most important variations in social data are due to trend, seasonal, and cyclical factors. Cyclical variations will be illustrated for both annual and monthly data. Table CII shows the method of computing cyclical variations for the mortality indexes:

TABLE CII

COMPUTATION OF CYCLICAL VARIATIONS FOR ANNUAL MORTALITY INDEXES CENTERED IN THE MIDDLE OF THE YEAR

Year (1)	Annual Mortality Index (2)	Annual Trend Values (3)	Ratio of Index to Trend Expressed as a Percentage (4)	Variations from 100, or Cycles (5)
1911	99.9	102.8	97.2	− 2.8
1912	102.3	102.9	99.4	− .6
1913	104.0	103.1	100.9	.9
1914	101.0	103.1	98.0	− 2.0
1915	98.3	103.1	95.4	− 4 6
1916	103.7	103.0	100.7	.7
1917	109.0	102.8	106.0	6.0
1918	122.7	102.4	119.8	19.8
1919	97.0	102.2	94.9	− 5.1
1920	104.6	101.8	102.8	2.8
1921	93.9	101.4	92.6	− 7.4
1922	92.9	101.1	91.9	− 8.1
1923	100.4	100.3	100.1	.1
1924	97.4	99.6	97.8	− 2.2
1925	97.1	98.9	98.2	− 1.8
1926	101.0	98.1	102.9	2.9
1927	95.7	97.2	98.5	− 1.5
1928	98.1	96.3	01.9	1.9
1929	98.3	95.2	103.2	3.2
1930	93.2	94.1	99.0	− 1.0

[10] Chapin, F. S., "Dependency Indexes for Minneapolis," *Social Forces*, Vol. V, No. 2, pp. 220-224.

The trend values are removed from the mortality indexes by taking the ratios of the original data to the trend values and expressing them as percentages. We have already referred to the trend values as the expected mortality indexes. We may also speak of these trend values as the normal death rate, or normal mortality index. Then, whatever the trend value is, it is 100.0 per cent, and the cyclical variations are determined by subtracting the ratio of original-data-to-trend-values from 100.0. If we take the trend as zero, we may express it as a straight line and indicate the cyclical variations graphically as follows (see p. 373).

The cyclical fluctuations above and below the line of trend are considerable. The 1918 rise above "normal" is greatest of any year and may be explained by the influenza epidemic which swept the country, but there is another factor involved. After 1915 the mortality index was going up. In 1916 it was almost 5 points higher than in the preceding year; this was the second year after the depression of 1914-15 began and may be due in part to the aftereffects of undernourishment and malnutrition during the period of depression. A similar change in the fluctuations occurred in 1923, about an equal length of time after the depression of 1920-21. We have previously noted that the trend in the mortality index, though parabolic in form, is downward; that would have to be explained by the interaction of a number of factors, such as improvement in medical care, rising economic standard of living, etc. The seasonal fluctuations are determined in part by weather conditions which favor the development of certain diseases and in part by other causes, such as lowered income in the winter months. While some of the same factors may be operating to determine the cyclical fluctuations, it will be seen that they operate in different ways and on different scales of magnitude; the element of accidental, or residual, causes enters into the cyclical conditions. We get a clearer picture of variations in the mortality index when it is analyzed into the three temporal forms.

Cyclical variations may also be measured by months. If this is done, an additional step in the computation is necessary to remove the seasonal factor from the monthly indexes. The monthly trend value must be estimated from the annual trend values, or the trend values must be computed on a monthly basis. In the illustration the trend values have been estimated from their annual change. At the middle of the year 1911 the trend value was 102.8, and at

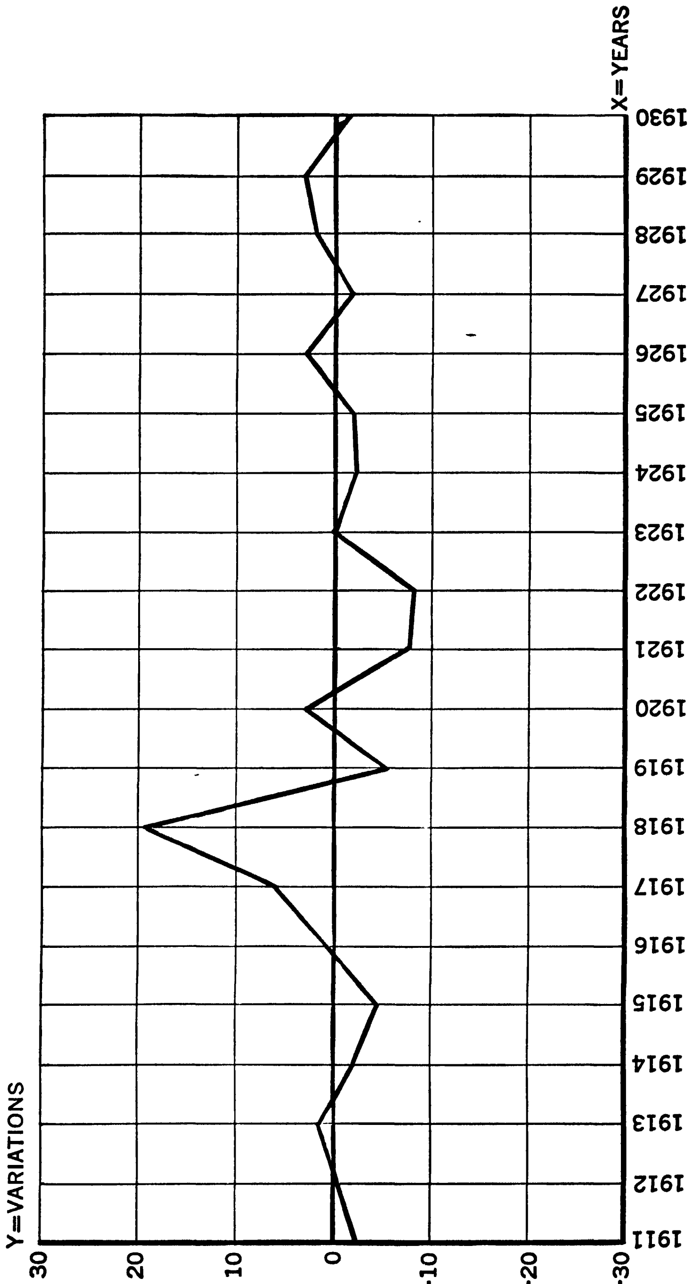


FIGURE LXVII.—CYCLES EXPRESSED AS DEVIATIONS FROM TREND—MORTALITY INDEXES

the middle of the year 1912 it was 102.9. That is, the rate is changing .1 per year; in later years it has changed as much as 1.1. If the latter figure had been used, it would have been necessary to show several changes in trend within the year. Since the trend values for the first two years have been used, the trend value for each month is assumed to be the trend value for the year. There is an assumption in assigning of the annual trend value to each month to which attention should be called: it is that the trend within the year is linear. While that is not strictly true, because the annual trend values are measured from a parabolic equation, the variation from linearity is so slight that it could not be indicated without using several decimal places which would suggest greater precision and reliability than the mathematical finesse warrants.

Table CIII presents the computation of cyclical variations by months in the mortality indexes for 1911 and 1912:

TABLE CIII

COMPUTATION OF CYCLICAL VARIATIONS OF THE MORTALITY INDEX BY MONTHS

Month (1)	Mortality Index (2)	Trend Values—Mortality Index (3)	Ratio of Index to Trend in Percentages (4)	Seasonal Index (5)	Cyclical Variations (4) − (5) (6)
1911					
January	114.6	102.8	111.5	112.4	− .9
February	113.0	102.8	109.9	109.3	.6
March	113.8	102.8	110.7	118.6	− 7.9
April	111.6	102.8	108.6	108.6	.0
May	94.5	102.8	91.9	97.0	− 5.1
June	85.6	102.8	83.3	87.0	− 3.7
July	102.4	102.8	99.6	91.9	7.7
August	92.9	102.8	90.4	92.0	− 1.6
September	87.3	102.8	84.9	92.0	− 7.1
October	93.7	102.8	91.2	94.8	− 3.6
November	91.2	102.8	88.6	93.2	− 4.6
December	97.7	102.8	92.1	103.2	−11.1
1912					
January	111.4	102.9	108.3	112.4	− 4.1
February	111.4	102.9	108.3	109.3	− 1.0
March	116.8	102.9	113.5	118.6	− 5.1
April	112.1	102.9	109.0	108.6	.4
May	91.1	102.9	88.6	97.0	− 8.4
June	84.7	102.9	82.4	87.0	− 4.6
July	97.5	102.9	94.8	91.9	2.9
August	100.0	102.9	97.2	92.0	5.2
September	100.8	102.9	98.0	92.0	6.0
October	99.2	102.9	96.5	94.8	2.3
November	94.3	102.9	91.7	93.2	− 1.5
December	108.1	102.9	105.1	103.2	1.9

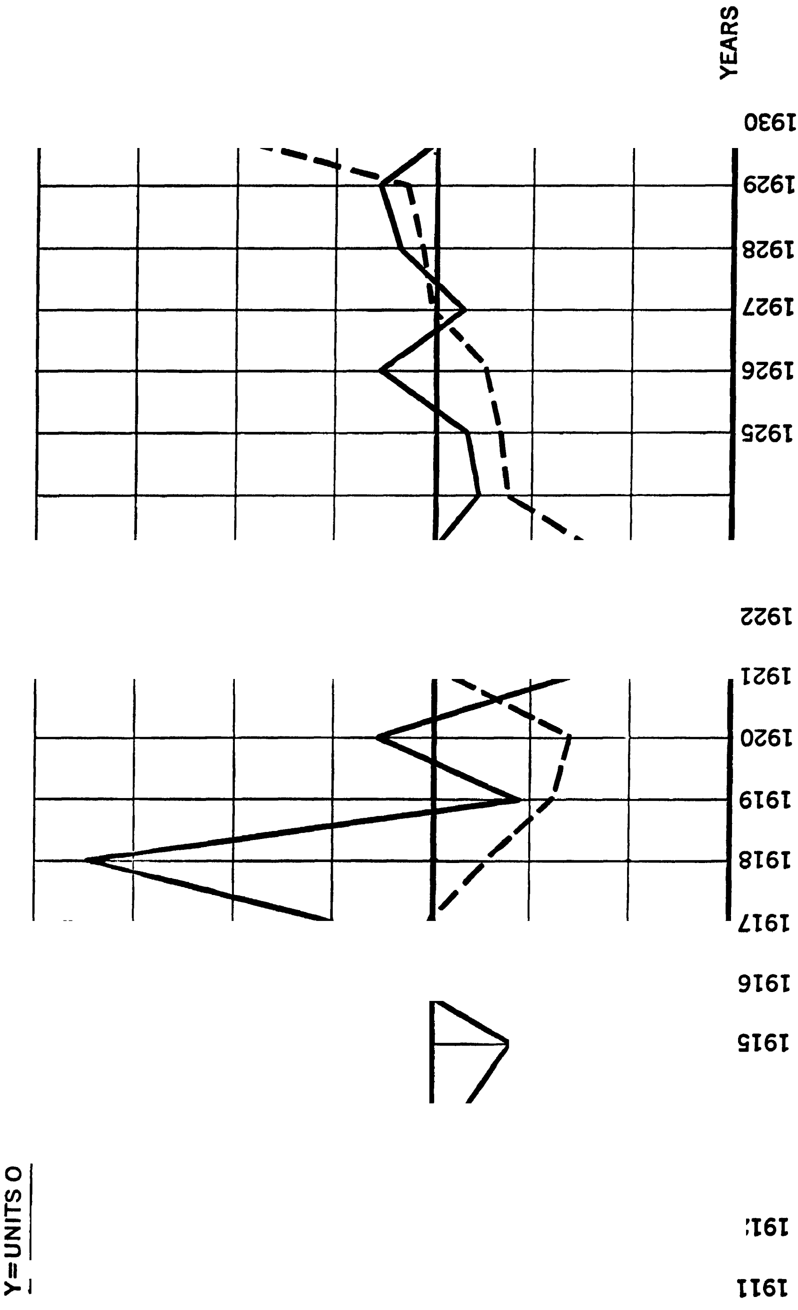
The monthly cyclical variations for other years would be determined in the same manner as those in this table. In Tables CII and CIII the cyclical variations were measured in units of the mortality index. If it is desirable to compare the cyclical variations of one social series with those of another, this cannot be done accurately when these variations are expressed in units of the variable. The difficulty can be overcome, however, by expressing the cyclical variations in terms of their respective standard deviations. After the computation of cycles this is a simple process. The cyclical variations are squared: the square root of the sum of the squares divided by the number of years, or months, equals the standard deviation of the cyclical variations. Then each cyclical variation is divided by the standard deviation. This will be illustrated by the cyclical variations of the mortality indexes and of poor relief in Indiana for the same years. Table CIV shows the process:

TABLE CIV

TRANSFORMATION OF CYCLICAL VARIATIONS IN UNITS OF THE VARIABLE TO UNITS OF STANDARD DEVIATION

Year	Mortality Indexes			Poor Relief Indexes		
	Cycles	Cycles Squared	Cycles in Units of σ—(2)÷5.70	Cycles	Cycles Squared	Cycles in Units of σ—(5)÷32.45
(1)	(2)	(3)	(4)	(5)	(6)	(7)
1911	− 2.8	7.84	− .49	− 1.5	2.25	− .04
1912	− .6	.36	− .11	7.7	59.29	.24
1913	.9	.81	.16	− 3.6	12.96	− .11
1914	− 2.0	4.00	− .35	42.3	1789.29	1.30
1915	− 4.6	21.16	− .81	64.7	4186.09	2.00
1916	.7	.49	.12	12.2	148.84	.37
1917	6.0	36.00	1.05	.5	.25	.02
1918	19.8	392.04	3.46	−14.1	198.81	− .44
1919	− 5.1	26.01	− .89	−37.5	1406.25	−1.15
1920	2.8	7.84	.49	−44.7	1998.09	−1.38
1921	− 7.4	54.76	−1.30	− 5.3	28.09	− .16
1922	− 8.1	65.61	−1.42	6.8	46.24	.21
1923	.1	.01	.02	−55.2	3047.04	−1.76
1924	− 2.2	4.84	− .38	−26.6	707.56	− .82
1925	− 1.8	3.24	− .32	−25.6	655.36	− .79
1926	2.9	8.41	.51	−17.1	292.41	− .53
1927	− 1.5	2.25	− .26	.3	.09	.01
1928	1.9	3.61	.33	3.4	11.56	.10
1929	3.2	10.24	.56	9.4	88.36	.29
1930	− 1.0	1.00	− .18	79.9	6384.01	2.46

The standard deviations have been computed from the data in columns (2) and (5). Columns (4) and (7) give the cyclical


YEARS
1930
1929
1928
1927
1926
1925
1922
1921
1920
1919
1918
1917
1916
1915
1911
Y=UNITS O

variations in units of standard deviation. These are seen to be much more nearly the same size than the units of the variables. To compare the cyclical variations more closely, columns (4) and (7) may be plotted. Figure LXVIII presents this comparison. The solid horizontal line represents zero deviation from the trend lines.

While there is some similarity between the variations of the two series, it is not close. The degree of similarity can be tested by the method of correlation.

5. CORRELATION OF TIME SERIES

The correlation of time series presents some special problems which do not appear when dealing with other types of distributions. The trend values and seasonal fluctuations of time series should not be treated by the method of correlation. The production of pig iron and the production of potatoes may both have upward trends, and a coefficient of correlation between the two series would perhaps be large, but it would be without significance because there is no reason to expect that these two series are functionally related. Seasonal fluctuations are related to specific conditions which affect particular series each year. If there is interdependence between two time series, it will be between the cyclical fluctuations. Consequently, before using the method of correlation for the study of time series the trend and the seasonal factor should be removed by methods already illustrated. A line of trend should be fitted to the data and a seasonal index computed. Then these variations may be subtracted from the original data, and the cyclical variations will be left. The usual methods of correlation may then be applied.

For purposes of illustration it is desirable to have two series of data which show marked correlation when the dependent variable is lagged, though it may show only slight correlation when the two variables are treated synchronously. For this illustration two series have been taken from Dr. Dorothy S. Thomas' study, made in England and Wales, of the relation of the business cycle to other social series.[11] The series are the business cycles and the phthisis, or tuberculosis, death rates which she computed for the years 1875 to 1894. Of the four periods studied, this shows the closest correlation of phthisis death rates, lagged two years, with the business cycle. The cycles in both cases are expressed as per-

[11] Thomas, D. S., *op. cit.*, pp. 187, 188, 197.

centage deviations of the annual items from secular trend, expressed in terms of standard deviation of each series. It will be remembered that the correlation for time series is to be computed for cycles only; since Dr. Thomas has already computed the cyclical variations of her data, it is not necessary here to repeat the process of determining these quantities. Table CV gives the first steps in computing the degree of correlation between the business cycle and the phthisis death rates when taken synchronously:[12]

TABLE CV

CORRELATION OF PHTHISIS DEATH RATES AND THE BUSINESS CYCLE, 1875 TO 1894, FOR ENGLAND AND WALES [1]

Year	Business Cycle—Deviations from Trend x	Phthisis Death Rates—Deviations from Trend y	x^2	y^2	yx-Products $-yx$	yx-Products yx
1875	.29	.67	.0841	.4489		.1943
1876	− .11	.31	.0121	.0961	.0341	
1877	− .34	.00	.1156	.0000	1.0807	
1878	−1.07	1.01	1.1449	1.0201	.5301	
1879	−1.71	.31	2.9241	.0961	.4794	
1880	.34	−1.41	.1156	1.9881	.7755	
1881	.55	−1.41	.3025	1.9881	.3910	
1882	.85	− .46	.7225	.2116		.8800
1883	1.60	.55	2.5600	.3025	.0000	.1224
1884	.36	.34	.1296	.1156	.2142	
1885	− .61	.00	.3721	.0000		.6875
1886	−1.19	.18	1.4161	.0324	.1200	
1887	− .55	−1.25	.3025	1.5625	.7790	
1888	08	−1.50	.0064	2.2500		1.9686
1889	.82	− .95	.6724	.9025		.6472
1890	1.02	1.93	1.0404	3.7249		.3182
1891	.64	.98	.4096	.9604		.0000
1892	− .37	− .86	.1369	.7396		
1893	−1.57	.00	2.4649	.0000		
1894	−1.02	−1.32	1.0404	1.7424		1.3464
	−8.54	−9.16	15.9727	18.1818	−4.4040	6.1646
	6.55	6.28				
	−1.99					

[1] Thomas, *op. cit.*, pp. 187, 197.

[12] For a comparison of the degrees of correlation found by using a variety of lags for social data, when correlated with an economic index, see Hexter, Maurice B., *Social Consequences of Business Cycles,* especially Chap. VIII Boston: Houghton Mifflin Co., 1925.

$$r =$$

$$c_x = \frac{-1.99}{20}$$

$$c_x^2 = .0100$$

$$-2.88$$

$$c_y^2 = .0196$$

$$\sigma_x = \sqrt{\frac{15.9727}{20} - .0100} =$$

$$\sigma_y = \sqrt{\frac{18.1818}{20} - .0196} = .943$$

$$r = \frac{\dfrac{6.1646 - 4.4040}{20} - .0140}{(.888)\,(.943)}$$

$$= + .105$$

This coefficient is quite low; it suggests that, if the phthisis death rate is correlated with changes in the business cycle, the effect is not synchronous with the change in the business cycle. When it is suspected that the changes in the dependent variable may occur later than changes in the independent variable, experiment with various lags is indicated. For purposes of illustration here, however, only the two-year lag of the phthisis death rate will be used. It has been found by Dr. Thomas that significant correlation exists between the business cycle and the phthisis death rate, when the latter is lagged two years. Table CVI gives the first steps in the computation of this coefficient of correlation.

The business cycles from 1875 to 1892 are used, and the phthisis cycles from 1877 to 1894. That is, when we speak of lagging the phthisis rate two years, we mean that the business cycle for 1875 is correlated with the phthisis rate of 1877 and so on throughout the 20-year period. The substitution in the formulas is identical with the substitutions shown for the data above without lag.

TABLE CVI

CORRELATION OF PHTHISIS DEATH RATES AND THE BUSINESS CYCLE, 1875 TO 1894, FOR ENGLAND AND WALES—PHTHISIS DEATH RATES LAGGED TWO YEARS

Year	Business Cycle—Deviations from Trend	Phthisis Death Rates—Deviations from Trend—Lagged Two Years			yx-Products	
	x	y	x^2	y^2	$-yx$	yx
1875	.29	.00	.0841	.0000	.0000	
1876	− .11	1.01	.0121	1.0201	.1111	
1877	− .34	.31	.1156	.0961	.1054	
1878	−1.07	−1.41	1.1449	1.9881		1.5087
1879	−1.71	−1.41	2.9241	1.9881		2.4111
1880	.34	− .46	.1156	.2116	.1564	
1881	.55	.55	.3025	.3025		.3025
1882	.85	.34	.7225	.1156		.2890
1883	1.60	.00	2.5600	.0000		.0000
1884	.36	.18	.1296	.0324		.0648
1885	− .61	−1.25	.3721	1.5625		.7625
1886	−1.19	−1.50	1.4161	2.2500		1.7850
1887	− .55	− .95	.3025	.9025		.5225
1888	.08	1.93	.0064	3.7249		.1544
1889	.82	.98	.6724	.9604		.7836
1890	1.02	− .86	1.0404	.7396	.8772	.7836
1891	.64	.00	.4096	.0000	.0000	
1892	− .37	−1.32	.1369	1.7424		.4884
	6.55	−9.16	12.4674	17.6368	−1.2501	9.0725
	−5.95	5.30				
	.60	−3.86				

$$r = \quad N \quad - c_x c_y$$

$$\frac{\dfrac{9.0725 - 1.2501}{18} - (.03)(-.21)}{(.832)(.967)}$$

$$= + .548$$

This coefficient is moderately high. It suggests that the effects of a change in the business cycle upon the phthisis death rate are considerable, two years after the change in the business cycle.[13]

Attention should be called to the fact that the probable errors of the two preceding coefficients of correlation have not been computed. Hitherto we have dealt with the correlation of frequency

[13] Both the above coefficients of correlation differ slightly from those Dr. Thomas published. The differences are doubtless due to minor variations in procedure.

distributions, where random sampling was assumed and where there was also assumed to be no relation between individual items of a single series. The situation is different in time series. Writing on this subject, Professor Warren M. Persons says: "There is a special objection to the application of the theory of probability to the particular economic data [time series] which constitute our material. If the theory of probability is to apply to our data, not merely the series but the individual items of the series must be a random selection. In fact, a group of successive items with a characteristic conformation constitutes our material. Since the individual items are not independent, the probable errors of the constants [such as coefficients of correlation] of a time series, computed according to the usual formulas, do not have their usual mathematical meaning. . . . Granting as one must that consecutive items of a statistical series are, in fact, related makes inapplicable the mathematical theory of probability."[14] Persons goes on to say that actually we do not know what, if any, meaning probable errors in time series would have. For that reason it is best not to compute them until some satisfactory method of calculating the range of variation is found.

Of course, the question of fitting a line of trend always arises in the correlation of cyclical fluctuations. The business cycles used here are the averages for several series of economic data used by Dr. Thomas: she used third degree parabolas for some of them and straight lines for others. The trend line for the phthisis death rate is a third degree parabola. Perhaps for beginning students the simplest line of trend, and the most flexible, is the moving average, unless it is fairly obvious that a straight line or a simple parabola will fit the data. But an exceedingly good case can be made out for the use of the moving average.[15] For preliminary purposes a freehand curve may be drawn through the plotted data; this is a rough guess at the trend.

6. EXERCISES

1. The following table gives the number of active cases carried by the Indianapolis Family Welfare Society from 1916 to 1931:

[14] Persons, W. M., "Some Fundamental Concepts of Statistics," *Jour. Amer. Stat. Ass'n,* Vol. XIX, New Series No. 145, March, 1924, p. 7.

[15] See Macaulay, Frederick R., *The Smoothing of Time Series.* New York: National Bureau of Economic Research, 1931.

TABLE CVII

ACTIVE CASES OF THE INDIANAPOLIS FAMILY WELFARE SOCIETY BY YEARS

Year	Cases	Year	Cases
1916	1028	1924	3227
1917	1446	1925	2638
1918	1534	1926	3048
1919	1474	1927	3872
1920	1306	1928	3690
1921	2501	1929	3106
1922	3605	1930	3997
1923	2499	1931	6169

(a) Fit a straight line to these data; a logarithmic curve; a second degree parabola; a four-year moving average.

(b) Find the mean deviation of the original data from each line of trend. Which shows the smallest mean deviation?

Fit a curve to the growth of population of the United States.

TABLE CVIII

POPULATION OF THE UNITED STATES AT EACH CENSUS, 1790 TO

Year	Population	Year	Population
1790	3,929,214	1870	38,558,371
1800	5,308,483	1880	50,155,783
1810	7,239,881	1890	62,947,714
1820	9,638,453	1900	75,994,575
1830	12,866,020	1910	91,972,266
1840	17,069,453	1920	105,710,620
1850	23,191,876	1930	122,775,046
1860	31,443,321		

The following table gives the active case load of the Indianapolis Family Welfare Society from 1924 to 1931 by months:

TABLE CIX

Month	1924	1925	1926	1927		1929	1930	1931
January	1329	847	1314	1673	1469	1528	2096	3450
February	1323	846	1230	1550	1459	1497	1992	3627
March	1088	850	1376	1440	1368	1371	1904	3518
April	760	745	981	1285	1147	1099	1678	3052
May	632	651	877	1086	983	931	1444	2335
June	582	769	853	1034	978	853	1284	1682
July	542	718	788	983	875	885	1166	1508
August	548	694	749	985	823	840	1155	1492
September	500	598	732	1041	868	845	1100	1632
October	610	719	750	1013	911	903	1301	2038
November	698	842	1127	1163	1084	1360	1903	2495
December	822	1209	1537	1412	1318	1941	2951	3274

(a) Compute seasonal indexes for the Family Welfare Data by the three methods discussed in this chapter.
(b) Compare the three indexes. Which seems best? Why? How would you use these indexes in planning a budget and employing personnel?
(c) Compute seasonal indexes of dependency for your own city or state.

4. Cyclical Variations:
(a) Determine the cyclical variations for the data in Table CVII.
(b) Determine the cyclical variations for the data in Table CIX.
(c) Compare these with some index of general business, for which corrections have been made for trend and seasonal variations. Are the variations similar? Does one series lag behind the other?

5. Correlation of time series:
(a) Compute the degree of correlation between the cyclical variations found in Exercise 4(a) and the cyclical variations in the index of general business which you used.
(b) Lag the relief case load by one year and compute the degree of correlation. Is there any significant difference between the size or sign of the two coefficients?
(c) Take two other time series, suggested by the instructor, that are believed to be related and compute the degree of correlation between the cyclical variations. This will be more interesting if the data used are local.

7. REFERENCES

Chaddock, R. E., *Principles and Methods of Statistics*, Chap. XIII.

Chapin, F. S., "Dependency Indexes for Minneapolis," *Social Forces*, Vol. V, No. 2, pp. 215-224.

Falkner, H. D., "The Measurement of Seasonal Variation," *Jour. Amer. Stat. Ass'n*, Vol. XIX, No. 146, pp. 167-179.

Hall, Lincoln W., "Seasonal Variation as a Relative of Secular Trend," *Jour. Amer. Stat. Ass'n*, Vol. XIX, No. 146, pp. 156-166.

Macaulay, F. R., *The Smoothing of Time Series*.

Mills, F. C., *Statistical Methods*, Chaps. VII, VIII, XI.

Thomas, D. S., *Social Aspects of the Business Cycle*, Chaps. I and II and Appendix A.

CHAPTER XIV

Vital Statistics

I. THE SCOPE OF VITAL STATISTICS

IN MOST extant books on statistical methods the subject of vital statistics is not given separate treatment, but since the facts are of great importance in the study of social problems and since there are some specific methods applicable to them, it seems desirable in a book of this kind to give this branch of statistical methods special consideration. Many of the methods previously discussed may be applied to vital statistics, after special methods have been used to bring the analysis to a certain point. Average death rates or birth rates over a period of time or in different localities may be desired; dispersions may be determined, index numbers computed, and correlations calculated. But in most cases some preliminary work should be done on the vital statistics before the application of these methods, and it is with this preliminary analysis that this chapter is mainly concerned.

What kinds of data may be called vital statistics? This question has sometimes been answered narrowly for administrative purposes as statistics of births and deaths; but it may be answered more broadly to include almost any kind of non-social data referring to human beings. Sometimes marriages are included, though they are social as well as biological matters. Whipple arrives at a definition of "vital statistics" through an analysis of the different divisions of demography. These divisions, he says, are genealogy, human eugenics, the census of population, registration of vital facts, vital statistics, biometrics, and pathometrics.[1] Vital statistics, according to Whipple, is the application of the statistical method to the study of vital facts, such as birth, marriage, divorce, sickness, and death. He omits the other divisions of demography. Pearl says, somewhat differently, " 'Vital statistics,' for which a

[1] Whipple, G. C., *Vital Statistics*, p. 2.

better term is *biostatistics*, is the special branch of biometry which concerns itself with the data and laws of human mortality, morbidity, natality, and demography."[2] These two definitions of vital statistics are not very harmonious, though they were given by two of the leading men who concern themselves with the types of data mentioned. Pearl regards vital statistics as a special branch of biometry but includes demography as a division of it. Whipple thinks of biometry as a division of demography coördinate with the division of vital statistics. Even if it were possible, it is unnecessary for our purposes to have a definition upon which everybody would agree. We shall simply take a few kinds of data usually regarded as vital statistics and illustrate methods of studying them. These types of facts are: population growth, marriages, births, deaths, and morbidity. Other types of facts which concern the social statistician might be included, but there is no doubt of the inclusion of any of these five.

2. POPULATION GROWTH

Population in a given geographical area increases because of births and immigration, and decreases because of deaths and emigration. The net result depends upon whether or not there is an excess of births and immigrants over deaths and emigrants. We are accustomed to expect an increasing population in all the great nations, but there are smaller areas in which population has declined and is declining. The statistician is concerned with both the quantity and the quality of changes in the population and with the possibility of forecasting future changes. It is much easier to measure past changes than it is to estimate changes that will take place in the future, but for many purposes it is desirable to make estimates with due allowance for a margin of error. The basis for estimating changes in population in the United States is the decennial census plus certain other data, such as births, deaths, immigration, and emigration. A rough way of estimating the population in intercensal years is the arithmetic method without reference to births, deaths, etc. For example, the population of continental United States in 1920 was 105,710,620 and in 1930 it was 122,775,046, which represents an increase of 17,064,426. If the population had increased the same amount each year, what would have been the population January 1, 1925? Since the time

[2] Pearl, Raymond, *Medical Biometry and Statistics*, p. 21. Philadelphia: W. B. Saunders Co., 1930.

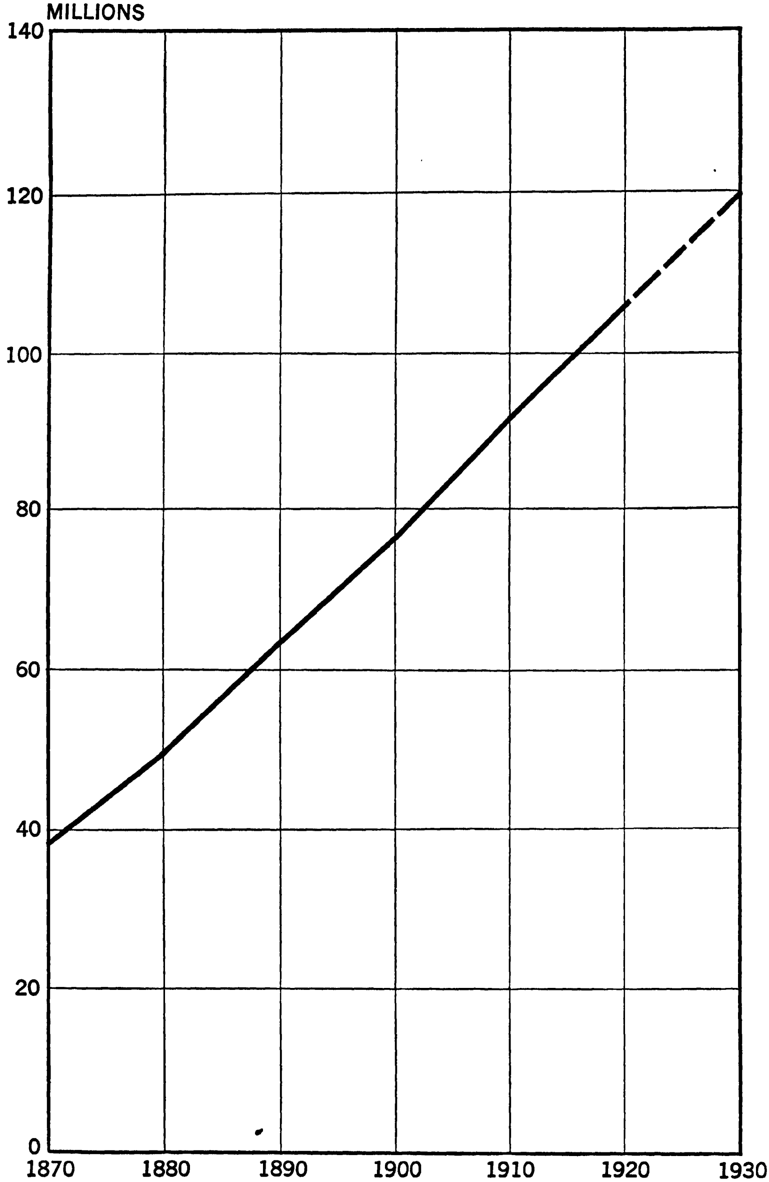


FIGURE LXIX.—ACTUAL POPULATION OF THE UNITED STATES, 1870-1920, AND PROJECTION OF THE CURVE TO 1930

between the census in 1920 and the census in 1930 was not 10 years but 10.25 years, we can divide the decennial increase by 10.25 to get the annual increase. The mean annual increase would be 1,664,822. If we multiply this figure by 5, we get 8,324,110 as the estimated increase, making a total estimated population of 114,034,730 for the country. Births, deaths, and migration have not been considered. This is the simplest way of estimating the population in an intercensal year but, of course, it is open to considerable error. By the same method we might assume that the rate of increase which obtained between 1920 and 1930 continued in 1931. We could then add 1,664,822 to 122,775,046 and get 124,439,868 for the population in 1931. But the longer this constant rate of increase is assumed, the larger the error is likely to be. Birth rates, death rates, and net increments or decrements from migrations change. Consequently, this arithmetic method is even approximately valid only for a short period of time, such as a decade. Allowance for births, deaths, and migration will be discussed below, when Dr. Whelpton's method of estimating population growth is considered.

Another way of estimating population change by the arithmetic method is to plot the census data to the natural scale and project the curve for future years. Figure LXIX shows the changing population of the United States from 1870 to 1920 and then projects the curve to 1930 to illustrate this method of estimation. If a freehand projection of the curve of population from 1920 to 1930 is drawn, the estimated population for 1930 is about 120,000,000, which is more than 2,500,000 less than the census. If the same increase in population occurred from 1920 to 1930 as from 1910 to 1920, and if this amount is added to the census of 1920, the estimated population in 1930 is 119,448,620, which is likely to be a more exact way of making the estimate than is the graph though in this case it happens to be less reliable. But in either case the error is considerable.

Where a large population is concerned, the geometric method of estimating population increase may be used. This method considers, not the absolute increase from one decade to another, but the percentage change. The formula for determining the rate of growth by the geometric method is as follows:

$$\log (1 + r) = \frac{\log P_1 - \log P_0}{N}$$

in which r is the annual rate of increase, P_1 the population at the end of the period, P_0 the population at the beginning of the period, and N is the number of years in the period. If the aim is to interpolate the population for intercensal years, the period chosen would be the decennium in which the interpolation is to be made. On the other hand, if the aim is to extrapolate (estimate population in future years) the population, we may use a period longer than 10 years so that the effect of long-time trend is more pronounced. For purposes of illustration we shall extrapolate the population for 1930, using 1870 to 1920 as the base period.

$$\log (1 + r) = \frac{\log 105{,}710{,}000 - \log 38{,}550{,}000}{49.5}$$

$$\log (1 + r) = \frac{8.024116 - 7.586115}{49.5}$$

$$= \frac{.438001}{49.5}$$

$$= .008848$$

$$1 + r = 1.02059$$

$$r = 1.02059 - 1 = .02059, \text{ or } 2.059 \text{ per cent per year increase.}$$

The census in 1870 was taken June 1, and in 1920 on January 1. So the period is 49.5 years. In 1930 the census was taken April 1. Hence the period from the 1870 census to the 1930 census was 59.75 years. The estimated population for 1930 would be 130,-295,017, or more than 7.5 millions more than it actually was. The rate of change had not been constant during this long period. Between 1870 and 1890 the rate of increase each year was near 3 per cent. Between 1910 and 1920 the annual percentage increase was about 1.5, and between 1920 and 1930 it was about 1.6. Hence, for extrapolation it is better to use the rate obtaining in the decade immediately preceding the period for which estimates are required. By the geometric method, using the period 1910 to 1920 as a base period, the estimated population, April 1, 1930, would be 122,771,705, which is 3,341 less than the census figure. This error is much less than the error involved in a 50-year base period. The geometric method may be used graphically also. Figure LXX shows the graphic method for the period 1870 to 1930. The graphic method is not useful for extrapolating the population, but it may be used for interpolating if only round numbers are

required. Figure LXX shows how the population in 1925 (note the broken lines) may be roughly estimated. It would be in the

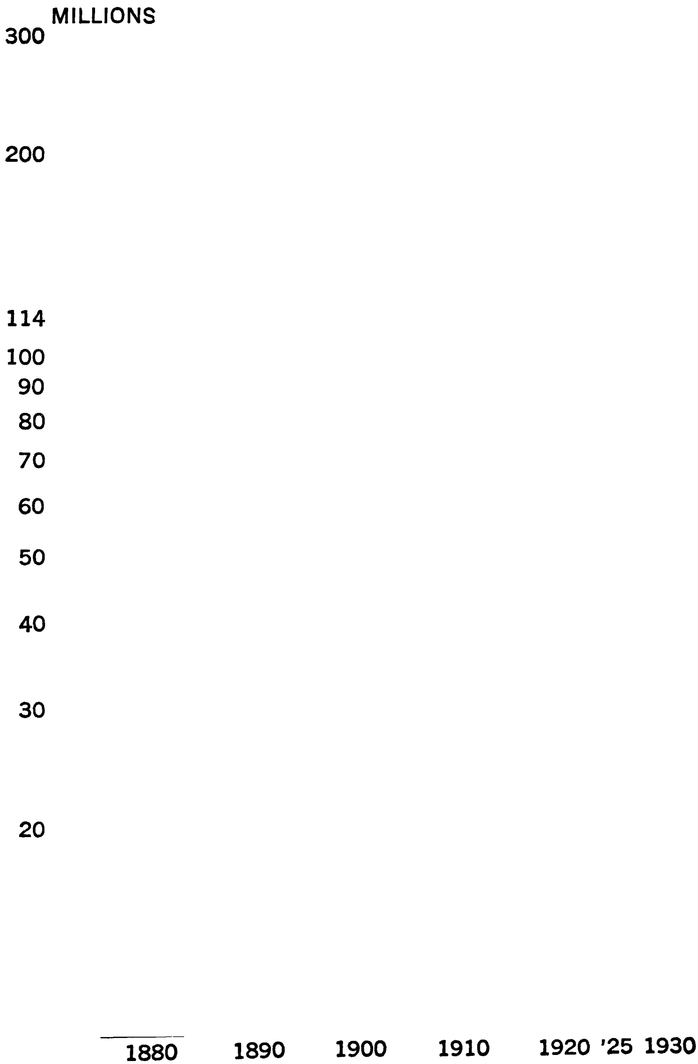


FIGURE LXX.—GROWTH OF THE POPULATION OF THE UNITED STATES

neighborhood of 114,000,000. If computations are to be based upon the estimated population, more exact methods are required.

Dr. P. K. Whelpton, of the Scripps Foundation for Population

Research, has presented a method of estimating population growth which involves only arithmetic, but it requires a great deal of detailed information not easily accessible to the average investigator using population data.[3] Dr. Whelpton bases his estimates upon survival rates for various age groups, birth rates of urban and rural white and negro considered separately, immigration rates, and rural-urban migration rates. He shows that there are good reasons for expecting a continuous decline in both the rate of increase and the numerical increase, and estimates that the population in 1975 will be about 175,120,000. There is a great deal to be said for this method of estimating population changes as against too much reliance upon more involved mathematical methods which make assumptions about the logistic nature of population growth. However, it is not a practical method for the student, because the special data are not readily available to him.

Sometimes it is desirable to know the population for a certain age, which is not given in the census returns. When the population is given only in age groups, a method of redistributing it according to the required age is, therefore, useful. This is done by means of a cumulative, or summation, curve. For instance, the 1930 population of Indianapolis by age groups was as follows:

TABLE CX

CENSUS OF INDIANAPOLIS BY AGE GROUPS[1]

Age Group in Years	Number by Age Group	Upper Limit of Age Group	Persons Less Than Upper Limit Age
Under 1	5,345	1	5,345
1–4	22,304	5	22,304
5–9	30,274	10	57,923
10–14	27,112	15	85,035
15–17	16,094	18	101,129
18–19	12,204	20	113,333
20–24	33,155	25	146,488
25–29	33,288	30	179,776
30–34	31,587	35	211,363
35–44	58,116	45	269,479
45–54	44,908	55	314,387
55–64	28,761	65	343,148
65–74	14,905	75	358,053
	358,905		358,905

[1] *Census of Indianapolis by Census Tracts*, Indianapolis Census Committee, 1931. Table I.

[3] Whelpton, P. K., "Population of the United States, 1925 to 1975," *Amer. Jour. Soc.*, Vol. XXXIV, No. 2, pp. 253-269.

Suppose it is desired to know the approximate number of the population who are 26 to 28 years of age. This age group could not be obtained from the reports of the census, but it can be estimated graphically as follows:

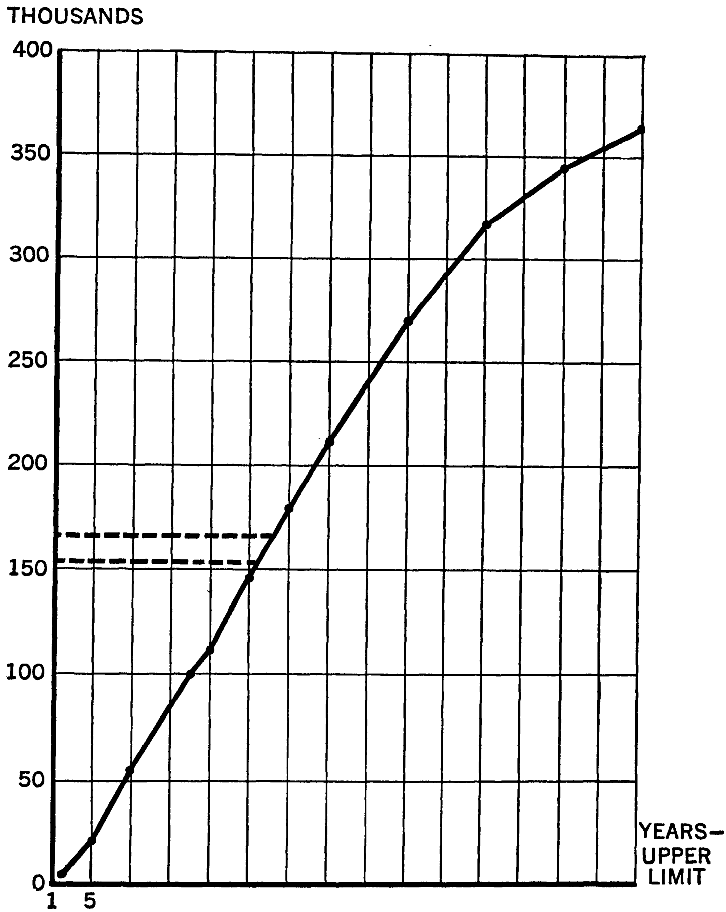


FIGURE LXXI.—CUMULATIVE CURVE OF INDIANAPOLIS POPULATION, 1930, AND ESTIMATION OF POPULATION 26 TO 28 YEARS OF AGE

The two broken parallel lines cut the curve at 26 and the upper limit of 28 years. In round numbers the population under 26 is 152,000 and under 29 is 173,000. The difference, 21,000, is the approximate population 26 to 28 years of age, inclusive. This is about as near as the population can be estimated, but it would be

satisfactory for some purposes, such as providing a base for computation of specific rates.

3. MARRIAGE AND DIVORCE RATES

Marriage rates may be computed in several ways, but the first problem is to define a marriage. A marriage is the union of a man and a woman in a given year or at any time whatsoever. Marriage rates may be rates of marriage within the year, or they may be the rates for all married persons in the population regardless of when they were married. A married person for the latter classification is anyone living with his or her spouse; widowed and divorced persons are not included.

The marital status of the population is often an important factor in the study of a variety of social problems. If the mean age at the time of marriage rises, one of the effects is to reduce the length of the childbearing period and, hence, the birth rate. Both the rate and the age of marriage vary with racial and national groups and with urban and rural populations, and the rate of marriage varies according to the ratio of men to women in the population, being highest when the ratio is considerably greater than 1.00. As the percentage of women gainfully employed increases, the percentage married appears to decrease. Death, crime, insanity, and pauper rates seem to be lower for married persons than for others. As divorce rates increase, there is a decrease in the social and biological significance of marriage. These observations indicate the importance of marriage to the work of the social statistician and emphasize to the student the importance of knowing how to compute marriage rates.[4]

The marriage rate in the United States for a given year would be the number of marriages consummated per 1,000 population over 15 years of age at the middle of the year. The rate of total marriage in the population at a given year is usually expressed as the percentage of the population over 15 years of age which is married.[5] The rates of total marriage may be refined by computing the percentage by sex and by age groups. If comparisons are to be made between years or decades, or between different geographical areas, this refinement is important because a peculiar variation in an age

[4] For a comprehensive statistical study of marriage, from which the above statements are derived, see Groves, E. R., and Ogburn, W. F., *American Marriage and Family Relationships,* especially Chaps. X-XVII, XIX. New York: Henry Holt & Co., 1928.

[5] Groves and Ogburn, *op. cit.,* Chap. XI.

group or in the sex ratio may explain differences in the total marriage rates.

Divorce rates are likewise of two kinds: rates for the total number of married persons and for the number of marriages consummated in a given year. In the first instance, the rate is expressed as the number of divorced persons in the population per 1,000 married persons in the population, and, in the second instance, divorces are expressed as the ratio of marriages to divorces. The use to be made of the rates will determine which kind of rate should be used.

4. BIRTH RATES

Births are reported as births and stillbirths. Of course, a stillbirth is a birth, but for purposes of clarity in the statistics it has become the custom to report the two kinds separately. For this reason it may be assumed, unless known to be otherwise, that a published birth rate is concerned with live-births only.

The "crude birth rate" is the number of live-births per 1,000 population in the year or month for which the rate is computed. This is the usual kind of rate published, though for some purposes the "refined birth rate" is preferable. The refined birth rate is the number of births per 1,000 women 15 to 44 years of age; it may be refined still further by expressing the rate as the number of births per 1,000 married women between the ages of 15 and 44. The trend in birth rates in the United States from 1919 to 1928 is shown in the following table:

TABLE CXI

BIRTH RATES, EXCLUDING STILLBIRTHS, IN THE REGISTRATION AREA OF THE UNITED STATES[1]

Year	Rate per 1,000 Population	Year	Rate per 1,000 Population
1919	22.3	1924	22.6
1920	23.7	1925	21.4
1921	24.3	1926	20.6
1922	22.5	1927	20.6
1923	22.4	1928	19.7

[1] *Birth Statistics*, United States Census, 1928, p. 4. Rates are based upon reports from the official Registration Area which includes all states except Nevada, New Mexico, South Dakota, and Texas.

When birth rates are computed by months, they are expressed as if the rate for each month were an annual rate. For example, if

the number of births in a city in the month of January is 1,000, this number is divided by the number of thousands of population, or women 15 to 44 years of age, and the result is multiplied by $\frac{365}{31}$, which gives a rate in terms of a year. The denominator of the fraction is always the number of days in the month for which the births are reported. The numerator is 366 in leap years.

Refined birth rates touch upon another matter which is omitted entirely by the crude rate, and that is fecundity. Fecundity is the productivity of women in terms of the number of children born, or, in another sense, it is the physiological capacity of women to conceive. If fecundity is thought of as actual productivity, then a rate is obtained by dividing the number of births by the number of thousands of women 15 to 44 years of age or by the number of married women within those ages. It is the latter to which Whipple refers in his discussion of fecundity.[6] Fecundity rates in this sense can be refined by computing fecundity by age groups. Such calculations are important in estimating population by Whelpton's method. But fecundity in the other sense referred to is less easy to measure. What proportion of women are sterile? What proportion of men are sterile? The fact that a couple does not have any children is not a satisfactory basis for the assumption of sterility in one or both married partners. The use of contraceptive methods accounts for the childlessness of some couples. Because of the difficulty of determining physiological fecundity in any large number of persons, reliable rates cannot be computed at the present time.

It will be noticed in Table CXI that the birth rate has been declining in recent years. That appears to be a general phenomenon in all Western countries. A trend line could easily be fitted to these rates by methods previously discussed and illustrated. We may also ask whether birth rates show cyclical and seasonal variations. As Thomas has shown, there are slight cyclical variations which are correlated with the business cycle.[7] These may be computed by the usual method of determining cyclical changes. There is little evidence to warrant belief that birth rates have marked seasonal variations, at least in the United States.[8]

[6] *Op. cit.*, pp. 247-249.
[7] See Thomas, D. S., *op. cit.*, Chap. IV.
[8] See White, R. Clyde, "The Human Pairing Season in America," *Amer. Jour. of Soc.*, Vol. XXXII, No. 5, pp. 800-805.

5. DEATH RATES

Death rates constitute an important part of vital statistics. There are general death rates based upon the number of deaths per 1,000 population and specific death rates for age groups and particular diseases. The latter may be based upon the number of deaths per 100,000 population or upon the number of deaths per 1,000 persons in the particular group.

The general, or crude, death rate is the one with which most people are familiar, and yet it has serious limitations for comparative purposes. It is fairly satisfactory for comparing the death rates at different times for the same area, provided the age and sex constitution of the population remain reasonably constant. On the other hand, when comparisons are made between general death rates for different areas, it is always an open question whether or not the rates are comparable on account of the possibility of important differences in age and sex constitution. Table CXII gives the annual general death rates for the registration area of the United States from 1919 to 1928, inclusive:

TABLE CXII

GENERAL DEATH RATES FOR THE REGISTRATION AREA OF THE UNITED STATES, 1919 TO 1928[1]

Year	Rate per 1,000 Population	Year	Rate per 1,000 Population
1919	12.9	1924	11.8
1920	13.1	1925	11.8
1921	11.6	1926	12.2
1922	11.8	1927	11.4
1923	12.3	1928	12.0

[1] *Mortality Statistics*, United States Census, 1928. All states except Nevada, New Mexico, South Dakota, and Texas are in the Registration Area.

During this period of 10 years the age and sex constitution of the population changed some, but the chances are that, if the population were standardized for these two factors, no great change would be apparent in the rates. However, it would be inaccurate to compare these rates with the general death rates for a particular state or with New England. The age and sex constitution for the state of Washington would be quite different from that of the country as a whole, and it would differ markedly from that of New England. Some method must be found to obtain a "corrected death

rate." This will involve using specific death rates for age groups and then combining them into a general rate.

The principle of the standard million of population must be introduced to compute the corrected death rate. The next table, from Pearl, gives the distribution of a standard million of population, both sexes together:

TABLE CXIII

STANDARD MILLION OF ACTUAL LIVING PERSONS (BOTH SEXES) IN THE UNITED STATES,

Age in Years	Persons per Million in Age Group	Age in Years	Persons per Million in Age Group
0– 4	115,806	55– 59	30,358
5– 9	106,321	60– 64	24,696
10–14	99,203	65– 69	18,294
15–19	98,728	70– 74	12,132
20–24	98,656	75– 79	7,269
25–29	89,104	80– 84	3,505
30–34	75,947	85– 89	1,338
35–39	69,672	90– 94	365
40–44	57,314	95– 99	80
45–49	48,682	100–104	39
50–54	42,491		

[1] Pearl, Raymond, *Medical Biometry and Statistics*, p. 262.

The formula used by Pearl to compute the corrected death rate is:

$$R_{C0} = 1000 \frac{\Sigma(L_x)(R_{sx})}{\Sigma(L_x)}$$

in which

R_{C0} = a corrected death rate
L_x = the number of persons of age x in the standard population
R_{sx} = the specific death rate at age x observed in the particular locality for which the corrected rate is being calculated

Before this formula can be applied, the specific death rates for different age groups in the particular locality under consideration must be computed. The equation for such specific death rates is:[9]

$$R_s = 1000 \frac{D_e}{E}$$

in which

R_s = specific death rate
D_e = deaths in a specified class of the population
E = number exposed to risk of dying, in the same specified class of the population from which the deaths come—age, sex, etc., might be basis of exposure

[9] Pearl, *op. cit.*, p. 212.

It will be seen that this equation can be used to compute the specific death rate for infants, for tuberculous patients, or for puerperal septicemia. We shall make use of it for computing specific death rates for age groups as a means of arriving at the corrected general death rate. The specific death rates for Indianapolis from September 1, 1930, to August 31, 1931, are computed in Table CXIV:

TABLE CXIV

Specific Death Rates in Indianapolis, September 1, 1930, to August 31, 1931

Age Group (1)	Deaths (2)	Population (3)	Specific Death Rates (4)
0– 4	709	27,649	25.6
5– 9	91	30,274	3.0
10–14	47	27,112	1.7
15–19	81	28,298	2.9
20–24	132	33,155	4.0
25–29	145	33,288	4.4
30–34	199	31,587	6.3
35–44	423	58,116	7.3
45–54	646	44,908	14.4
55–64	828	28,761	28.8
65–74	896	14,905	61.1
75 and over	866	5,683	152.4
	3,953	363,736	

These specific death rates will now be used to compute a corrected death rate for the city of Indianapolis. Table CXV gives the computations required for the equation (see following page).

The totals in columns (2) and (4) will now be used in the equation:

$$R_{C0} = 1000 \frac{12607.8}{1000000}$$
$$= 12.6$$

This figure, 12.6, is the corrected death rate for Indianapolis, that is, it is the death rate Indianapolis would have had if the city had the same population distribution the country as a whole had in 1910. This rate can now be compared with a corrected death rate for any other city of the country.

Attention may be called to the fact that a corrected death rate is a weighted average of the local specific death rates. The weights

TABLE CXV

EXPECTED DEATHS IN INDIANAPOLIS, SEPTEMBER 1, 1930, TO AUGUST 31, 1931

Age Group (1)	Persons in Actual Population per Million, in Thousands (2)	Specific Death Rates (3)	(2) × (3) (4)
0– 4	115.806	25.6	2964.6
5– 9	106.321	3.0	319.0
10–14	99.203	1.7	168.6
15–19	98.728	2.9	286.3
20–24	98.656	4.0	394.6
25–29	89.104	4.4	392.1
30–34	75.947	6.3	478.5
35–44	126.986	7.3	927.0
45–54	91.173	14.4	1312.9
55–64	55.054	28.8	1585.6
65–74	30.426	61.1	1859.0
75 and over	12.596	152.4	1919.6
			12607.8

consist of the proportions of the population in each age group of the standard million of population.[10]

Two other kinds of corrections may be made in the computation of death rates: Some persons die in a locality who do not live there; for example, at a large general hospital which serves no definite geographical area. Some persons who live in the community die away from the locality. Should consideration be given to these facts, or can we assume that as many non-residents will die in the city as residents die away from the city? An exact death rate would have to take these questions into consideration. It might happen that a city had particularly elaborate hospital facilities and that more non-residents would die in the city and be reported to the local authorities than the number of residents dying away from the city. In the Indianapolis data used above only persons who had a residence in the city were used. No check could be obtained on those who died away from the city. Consequently, both specific and general death rates are lower than they should be. For ordinary purposes, it may be assumed that the residents dying away from home and the non-residents dying in the city are equal; for more exact calculations, their equality or inequality should be determined if possible.

[10] See Pearl, *op. cit.*, pp. 171-174.

It is well known that seasonal variations occur in death rates, that there are cyclical variations, and that over a long period of time a secular trend is perceptible. These measures may be determined after the manner described and illustrated in Chapter XIII.

6. MORBIDITY

Social statisticians, as well as public health officials, are interested in sickness, or morbidity. They would like to know the case rates in the population for many particular diseases, but, because sickness is so generally regarded as a personal matter, reliable data on the prevalence of disease are almost nil. This is not so true of what are known as "reportable diseases," that is, infectious diseases which the attending physician is required by law to report to some central health agency. Even some of the infectious diseases, for example, gonorrhea and syphilis, are not reported regularly because the physician regards his relation to his patient as personal and confidential, and declines to list his private patient among those having certain infectious diseases with which a social stigma is associated. The United States Public Health Service gets weekly reports from American consuls for the following diseases: cerebrospinal meningitis (epidemic); cholera, Asiatic; cholera nostras, cholerine, or gastroenteritis; diphtheria; measles; plague, human; plague, rodent; poliomyelitis (acute anterior poliomyelitis or infantile paralysis); scarlet fever; smallpox; tuberculosis; typhoid fever (enteric fever, typhus abdominalis); typhus fever (typhus exanthematicus); and yellow fever.[11] Similar reports are received from local health officials within the United States for chicken pox, diphtheria (carriers not included), influenza, measles, mumps, pneumonia (all forms), scarlet fever, smallpox, tuberculosis (all forms), typhoid fever, whooping cough, cerebrospinal fever, dengue, lethargic encephalitis, pellagra, poliomyelitis (infantile paralysis), rabies (in man) (developed cases), rabies (in animals), typhus fever.[12] The non-reportable diseases, which are non-infectious or only slightly so, are not reported to health agencies with sufficient completeness to make the data reliable. State and city health departments often try to get these diseases reported, but there is no way of determining what

[11] *Public Health Reports, United States Public Health Service,* February 6, 1931, p. 285.
[12] *Ibid.*, p. 286.

percentage of the total cases are reported. To obtain adequate statistics of morbidity, for both infectious and non-infectious diseases, is a problem of health organization and of persuasion of the medical profession of the public interest at stake in all forms of disease.

Because of the lack of adequate morbidity statistics, the vital statistician and the student of social problems are strongly tempted to assume that there is a constant ratio between the number of deaths from a specific disease and the total number of cases of the disease. If this were a fact, the number of cases could be inferred from the ratio of mortality to morbidity, but this is unreliable. Discussing this question, Pearl says: "Mortality is not and never can be a good index of morbidity, generally speaking. What actually is done is to weaken and impair the value of the statistics for the study of *mortality* in the hope to make them a little better indices of *morbidity*. . . . It is thought desirable to get as complete records as possible of the *prevalence* of cancer in the population, as a disease. Therefore, the rule is that, in general, if a person dies who is known to have had cancer prior to death, the death is charged to cancer. In consequence, it results that no one can get from the official statistics an accurate answer to the question: 'How many persons per 1000 living did cancer kill in 1920?' Instead, what he gets is information as to how many persons died per 1000 living in 1920, who had cancer before they died, assuming that the diagnosis is correctly made in every case. The latter information, as anyone with a logical mind will at once perceive, is quite different from the former."[13]

If morbidity rates are to be computed, fully understanding that they are open to wide margins of error, they may be crude or specific rates. If they are crude rates, then the number of cases per 100,000 population is the usual measure for specific diseases. Specific case rates would be determined from the number of cases per 1,000 persons belonging to the class exposed—e.g., age group, sex, etc. Under any circumstances the results warrant only very limited confidence.

7. EXERCISES

1. The following table gives the population of New York City from 1900 to 1930 inclusive:

[13] *Op. cit.*, p. 103.

TABLE CXVI

POPULATION OF NEW YORK CITY, 1900 TO 1930

Year	Population	Year	Population
1900	3,437,202	1920	5,620,048
1910	4,766,883	1930	6,930,446

(a) Compute the rate of growth of population in each decennium and for the 30-year period by the arithmetic method.
(b) Compute the rate of growth of population in each decennium and for the 30-year period by the geometric method.
(c) Compare the results obtained from using the arithmetic and geometric methods.
(d) Estimate the population at each intercensal year between 1920 and 1930. Using the same basis of estimate, what would you expect the population to be in 1940?
NOTE: The census was taken on the following dates: 1900, June 1; 1910, April 15; 1920, January 1; 1930, April 1.

2. Table CXVII gives the number of persons out of work in Illinois at the time of the United States Census in 1930:

TABLE CXVII

PERSONS OUT OF A JOB, ABLE TO WORK, AND LOOKING FOR A JOB, CLASS A, ILLINOIS, APRIL, 1930[1]

Age Group	Number	Age Group	Number
10–14 years	84	45–49 years	21,237
15–19 years	23,205	50–54 years	16,758
20–24 years	36,447	55–59 years	12,548
25–29 years	27,808	60–64 years	9,186
30–34 years	23,490	65–69 years	5,494
35–39 years	24,678	70 and over	2,788
40–44 years	23,035	Unknown	241

[1] *Unemployment Bulletin, Illinois*, United States Bureau of the Census, 1931.

(a) Determine graphically the approximate number unemployed who are 26 but less than 29 years of age.
(b) Determine graphically the approximate number unemployed who are 46 but less than 48 years of age.

3. Table CXVIII gives the births by months in Indiana for 1928 to 1930, inclusive:

TABLE CXVIII

BIRTHS IN INDIANA,[1] 1928 TO 1930, BY MONTHS. POPULATION OF INDIANA: 1928, 3,176,000; 1929, 3,207,689; 1930, 3,238,000

Month	Births	Month	Births
1928		1929	
January	4,962	July	4,984
February	4,646	August	4,801
March	5,147	September	4,361
April	4,629	October	4,359
May	4,594	November	4,303
June	4,479	December	4,645
July	4,770	1930	
August	4,825	January	4,733
September	4,572	February	4,433
October	4,575	March	4,795
November	4,428	April	4,583
December	4,560	May	4,647
1929		June	4,544
January	4,552	July	4,996
February	4,443	August	4,992
March	5,095	September	4,565
April	4,352	October	4,446
May	4,384	November	4,241
June	4,506	December	4,300

[1] *Monthly Bulletin*, Indiana State Board of Health, January, 1928, to December, 1930.

(a) Find the crude birth rates for each month, expressed as annual rates, for the above data.

(b) For a seasonal index of births to be reliable more years are required than are given in this table, but this may be used for illustrative purposes. Compute the seasonal variations in births, if any.

(c) Plot the crude birth rates. Is there evidence of a cyclical decline following in the wake of the depression which began in 1929?

NOTE: The population data may be computed from the census reports.

4. Table CXIX gives the deaths in the United States from 1914 to 1928:

TABLE CXIX

DEATHS FROM ALL CAUSES IN THE UNITED STATES, 1914 TO 1928, AND THE ESTIMATED POPULATION OF THE REGISTRATION AREA[1]

Year	Estimated Population	Deaths
1914	65,813,315	898,059
1915	67,095,681	909,155
1916	71,349,162	1,001,921
1917	74,984,498	1,068,932
1918	81,333,675	1,471,367
1919	85,166,043	1,096,436
1920	87,486,713	1,142,558
1921	88,667,602	1,032,009
1922	93,241,643	1,101,863
1923	96,986,371	1,193,017
1924	99,200,298	1,173,990
1925	103,108,000	1,219,019
1926	105,167,000	1,285,927
1927	108,327,000	1,236,949
1928	114,495,000	1,378,675

[1] *Mortality Statistics*, United States Bureau of the Census, 1928.

(a) Compute the crude death rate for the United States from 1914 to 1928.
(b) Fit a line of trend to these rates.
(c) Compute the cyclical variations of these death rates.

5. Table CXX gives the number of deaths in the United States in five-year age-intervals for the year 1928:

TABLE CXX

DEATHS IN THE UNITED STATES IN FIVE-YEAR INTERVALS, 1928, AND THE ESTIMATED POPULATION IN EACH INTERVAL FOR THE REGISTRATION AREA[1]

Age Group	Estimated Population	Deaths
0– 4	12,479,955	216,090
5– 9	12,365,460	25,245
10–14	11,563,995	19,494
15–19	10,190,055	33,226
20–24	10,075,560	43,445
25–29	9,846,570	44,062
30–34	8,701,620	46,454
35–39	8,472,630	56,754
40–44	6,869,700	62,218
45–49	6,297,225	70,759
50–54	5,152,275	82,319
55–59	3,892,830	89,367

TABLE CXX—(*Continued*)

Age Group	Estimated Population	Deaths
60–64	3,205,860	101,676
65–69	2,289,900	117,229
70–74	1,488,435	118,904
75–79	915,960	107,293
80–84	457,980	78,343
85–89	114,495	43,173
90 and over	57,248	20,164
Unknown	114,495	2,460
Total	114,552,248[2]	1,378,675

[1] *Mortality Statistics*, United States Bureau of the Census, 1928.

[2] The total is a little higher than the estimate for the whole registration area, as given by the census, because the percentages in each age group have been carried to only one decimal place, and the percentage for the group 90 and over is only .05 per cent and is not given in the reports of the census. The population for this group has been estimated by the author.

(a) Compute a corrected death rate from the above data. How does it compare with the crude rate for 1928?

6. Obtain the mortality data for your own state, if available, and compute:

(a) The crude death rates from 1919 to 1928. How do they compare with the national rates?

(b) The corrected death rate for 1928. How does it compare with the national rate?

8. REFERENCES

Newsholme, Sir Arthur, *Elements of Vital Statistics*, New Edition.

Pearl, Raymond, *Medical Biometry and Statistics*, Chaps. VII-IX.

Thomas, Dorothy S., *Social Aspects of the Business Cycle*, Chaps. III-V.

Whipple, G. C., *Vital Statistics*, Chaps. IV-XII.

CHAPTER XV

Rating Scales

1. THE FUNCTION OF RATING SCALES

THE discussion and illustration of rating scales might have been given in Chapter V, "Collection and Assembling of Data," but there is a logical difference between the kind of data discussed in that chapter and the kind sought by means of a rating scale. The data discussed in Chapter V are obtained mainly by a counting scheme, whereas a rating scale is intended to show degrees of difference in a single variable. The methods of statistical analysis described in preceding chapters may be applied to data gathered by means of a rating scale, but the theory of the rating scale is of sufficient importance to justify treatment in a separate chapter.

During the past decade increasing emphasis has been placed upon measurement in psychology, education, and the social sciences. Many of the traits to be measured, however, present great technical problems for two reasons: first, because we are not in the habit of thinking of them in quantitative terms, and, second, because the invention of measuring sticks is difficult. However, experimentation has given some valuable, and perhaps more hopeful, results. Is a man a pacifist or a militarist, or does he seem to occupy a middle-of-the-road position with respect to these two popular conceptions? Is it possible to mark off degrees of attitude toward war, ranging from complete pacifism to complete militarism? Are all blind persons blind in the same degree, or should definite degrees of blindness be distinguished? Are there degrees of psychoneurotic personality, or is psychoneurotic personality a fixed and definite thing like pneumonia? The attitudes or conditions mentioned are in fact variables. But how are we to measure the degrees of variableness? That is the function of a rating scale. Scales of pacifism-militarism and of blindness must be devised. The scale will be analogous to the division of length into feet and

inches, and variations infinitely small may be indicated. That is, the assumption back of the rating scale is that attitudes and social conditions are continuous variables.

It may be objected that no accurate scale, comparable to linear measure, can be devised for attitudes and social conditions commonly assumed to be qualitative and, hence, not the proper objects of measurement. Until recently this viewpoint was generally accepted, but the introduction of statistical concepts into physics and chemistry has tended to change it. It has been seen that successive measurements of the same material object do not agree exactly, if the unit of measurement is made indefinitely small. These successive measurements tend to distribute themselves in a normal frequency curve. Consequently, it is logical to reason that, even if an attitude cannot be measured exactly by different people or by the same persons at different times, attempts at measurement are justified and normal errors are to be expected. We cannot say that one science is quantitative and that another is totally lacking in this characteristic. The physical, biological, and social sciences might themselves be arranged along a rating scale according to degrees of precision of measurement attainable in the present state of scientific technique. It is quite likely that the standard deviation of measurements of an attitude by means of a rating scale would be larger than the standard deviation of successive measures of the expansion of a piece of steel under specified temperatures, but, if the validity of the rating scale can be determined, the results are reliable within the range of ascertained error.

Experimentation with rating scales has proceeded far enough to reveal certain tests for reliability and validity. For convenience we may classify rating scales into sociometric and attitude scales. The sociometric scale is used for measuring aspects of social institutions, and the attitude scale is used for measuring the mental set of an individual toward a certain type of reaction. (Perhaps a third type should be mentioned, such as the test for degree of blindness given below, but this type is really a physiometric scale and is mentioned in this book only because of the social implications of blindness.) The first is concerned with material culture or physical conditions; the second is concerned with the reaction organization of an individual which has been built up in a cultural and physical environment.

Six tests for the reliability and validity of a sociometric scale

may be distinguished. (1) Reliability may be tested by having different observers rate the same subject. The degree of correlation between these ratings constitutes a measure of the reliability of the scale. Different persons should be able to rate the same subject similarly, if the scale is reliable.[1] (2) The scale must have general validity. That is, it must not be seriously affected in validity when it is applied to different subjects of the same class. If it is a home rating scale, it should be applicable to homes in any city or rural community. Furthermore, the degree of correlation between the results obtained on a given scale and on some other scale that has been standardized for the same class of subjects should be high, if the given scale is valid. (3) The scale must make possible the establishment of reasonable norms for the subjects to which it is to be applied. This norm will be statistical and will be represented by a curve of distribution which is *normal* for this class of objects; the curve may approach the form of a normal curve of error, or it may be skewed. But application to a random sample of the subjects should make possible the establishment of a norm. For such a distribution measures of central tendency and dispersion may be computed. (4) Factors which enter into the construction of the scale should be generally available to the investigator. Availability implies accessibility to the subjects to be studied and reasonably exact definition of terms. If the terms are ambiguous, the reliability and validity of the scale are likely to be low. (5) Assuming the availability of all factors of importance to the problem, the scale should take into consideration all significant aspects for which quantitative evaluations can be secured. Here the judgment of the worker is paramount. He may resort to experiment to determine what are the important factors and aspects of factors, or he may rely upon his own judgment and that of other qualified persons. For example, in constructing a scale for rating the living rooms of homes, all the objects which have diagnostic value should find a place in the scale. (6) Each factor in the scale should be weighted according to its relative significance. If we are trying to measure the importance of blindness as a social problem in a state by means of an index, it is of first importance to know what weight to attach to the blindness of a person who cannot distinguish light from darkness and to the blindness of a person who can walk around

[1] Lundberg, G. A., *op. cit.*, pp. 248-252. Quoted from Gould, K. M., *A Sociometric Scale for American Cities*, pp. 53-57. M. A. Thesis, Columbia University, 1921. Points (2) to (6) above are taken from this source.

but cannot read. Some standard of significance of factors must be set up; it is similar to the problem of weighting prices to obtain a price index. Weighting is an important problem in the establishment of the validity of a given scale. If different weights are used in the given scale and in some other standardized scale with which the given scale is to be compared, the results obtained and the degree of correlation discovered between results might be low because of the difference of weights. Consequently, weights also must be tested for validity.

Somewhat similar tests for reliability and validity may be used for attitude scales. Dr. Goodwin B. Watson has used the following tests of the validity of results obtained in the use of a test of "fair-mindedness":[2] (1) Examination of the tests with reference to what they seem to be measuring; (2) correlations between each form of the test and the test as a whole; (3) a study of the scores obtained by individuals who are selected by their group as most fair-minded; (4) individuals who are supposed, by those who know them well, to have pronounced lines of prejudice are given the test, and their reactions compared with those which would be anticipated; (5) certain groups who might be supposed to possess certain lines of prejudice are studied by the test and the result compared with the assumptions of competent judges as to the lines of prejudice that might be expected to exist within the given groups; (6) the tests are examined to determine to what extent they are measures of intelligence or opinion rather than of prejudice. Thurstone employed what he called objective tests of reliability and validity to his rating scale for attitude toward the church.[3] They are: (1) the probable error of the scale value, which is equal to the product of half the standard deviation of the scale values and the standard error; (2) the existence of ambiguous statements in the test is determined by the scale-distance, the X-value, between the first and third quartiles: if the distance is great, and the curve flat, the statement is ambiguous; (3) the existence of an irrelevant statement is determined by comparing the ratings on other statements of similar character; if the statement is relevant, the other ratings should be distributed in the form of a normal frequency curve. Thurstone's criteria are wholly

[2] Watson, G. B., *The Measurement of Fair-Mindedness*, p. 19. Teachers College, Columbia University, 1925.

[3] Thurstone, L. L., and Chave, E. J., *The Measurement of Attitude*, pp. 42-56. University of Chicago Press, 1929.

objective and should be applied to the results obtained from any attitude scale used.

Four types of rating scales will be used for purposes of illustration: (1) the scale for blindness developed by the Committee on Central Statistics of the Blind; (2) Chapin's Scale for Rating Living Room Equipment; (3) the Matthews revision of Woodworth's Psychoneurotic Inventory; and (4) Thurstone and Chave's scale for measuring attitudes toward the church.

2. A BLINDNESS SCALE

We are accustomed to think of blindness as a unitary term. A blind person is simply one who cannot see. But some persons who are classified as blind cannot distinguish light from darkness, while others are able to walk about unaided but cannot read. Obviously there are degrees of blindness. The Committee on Central Statistics of the Blind has taken the Snellen scale for measuring visual perception and has given the following descriptive terms to the five degrees of blindness recognized by Snellen: (1) totally blind or having "light perception only"; (2) having "motion perception" and "form perception"; (3) having "traveling sight"; (4) able to read large headlines; (5) "borderline" cases. For each of these classes Snellen has exact measurements of the amount of visual perception. The scale is reproduced on page 410.

This scale is interesting for two reasons as an illustration of quantitative analysis of a trait: first, because it makes clear the fact that blindness is a variable, and, second, because it shows the "rough tests for lay workers" in a column parallel to the Snellen measurements. To most people a blind person is simply a blind person; qualifications to suggest degree of blindness are not made. We see here a trait that may be measured exactly by means of the Snellen scale. The five divisions are chiefly for the "lay worker." Actually blindness exists in all gradations, however small, from total absence of visual perception to the so-called borderline cases. It is a continuous variable, and in a list of 1,000 blind persons who had been tested we should expect to find Snellen measures continuous from 0 to some point arbitrarily chosen as the maximum visual perception consistent with the definition of blindness. The rough test given for lay workers makes blindness a discontinuous variable for no other reason than that rough tests are guesses and not measures. The pigeonhole type of social classification is well

PROPOSED TABLE FOR UNIFORM GROUPING OF THE BLIND BY AMOUNT OF VISUAL PERCEPTION

Group	Description of Group	Snellen Measurements[1] of Visual Perception: At various distances (feet)	At a fixed distance (20 feet)	At a fixed distance (6 meters)	Rough Tests[1] for Lay Workers
1	Totally blind or having "light perception" only[2]	0 Up to but not including 2/200	0 Up to but not including 20/2000	0 Up to but not including 6/600	No vision, or light perception only[2] Up to but not including Perception of motion of hand at a distance of 3 feet (arm's length) or less
2	Having "motion perception" and "form perception"	2/200 Up to but not including 5/200	20/2000 Up to but not including 20/800	6/600 Up to but not including 6/240	Ability to perceive motion or form of hand at a distance of 3 feet (Arm's length) or less Up to but not including Ability to count fingers at a distance of 3 feet (arm's length)
3	Having "traveling sight"	5/200 Up to but not including 10/200	20/800 Up to but not including 20/400	6/240 Up to but not including 6/120	Ability to count fingers at a distance of 3 feet (arm's length) Up to but not including Ability to read large letters (such as newspaper headlines)
4	Able to read large headlines	10/200 Up to but not including 20/200	20/400 Up to but not including 20/200	6/120 Up to but not including 6/60	Ability to read large letters (such as newspaper headlines) Up to but not including Ability to read large print (larger than 14-point type)
5	"Borderline" cases[3]	20/200 or more but not sufficient for use in an occupation or activity for which eyesight is essential.	20/200 or more	6/60 or more	A. Ability to read 14-point type but not 10-point type. B. Ability to read 10-point type but with a defect of vision (such as limited field, etc.) so great as to be a marked handicap.

[1] All measurements and tests apply to vision in the better eye after correction.
[2] "Light perception" is defined to mean just sufficient vision to distinguish light from darkness.
[3] Examination by an eye physician is recommended for all cases but individuals in group 5 should not be finally classified except upon the basis of such an examination. Certain eye conditions such as high progressive myopia, greatly restricted field of vision, etc., may constitute such a severe handicap in activities for which eyesight is essential that even when the individual has a visual acuity of 20/200 or more, he is, for occupational purposes, blind. These are the "borderline" cases.

This classification has been drawn up by the Committee on Central Statistics of the Blind.

illustrated by the rough tests, and is sharply contrasted with the statistical conception of variability. Continued experimentation with rating scales should result in a gradual decrease in the use of the former and a gradual increase in the use of the latter.

The Snellen scale was developed by recording the visual perception of patients at varying distances. Visual perception is defined

in terms of linear measurement. Other rating scales will be defined in other terms, but the aim is to find measures that can be applied with reliability and to express the results in quantitative form.

3. CHAPIN'S SCALE FOR RATING LIVING ROOM EQUIPMENT

Sociologists have for a number of years been interested in developing some measure of homes. The rural sociologists have made housing surveys for the purpose of determining that part of the farmer's standard of living represented by the house he lives in. For the most part these surveys have not had the scientific value which it is desirable that they should have. The sociologist is interested in the home from the viewpoint of the standard of living of the family but perhaps more from the viewpoint of the home as a center of social interaction. Historically the concept of social interaction has been chiefly a descriptive term, but efforts are now being made to find some way of treating it as a variable in the statistical sense. In order to limit the size of his rating scale, Professor Chapin constructed a scale for the living room of the home only. He says: "The sociological assumptions underlying the Living Room Scale developed to measure socio-economic status as defined are: (1) the living room of a home is the room most likely to be the center of interaction of the family; (2) the living room equipment reflects the cultural acquisitions, the possessions, and the socio-economic status of the family."[4] Here is an illustration of the necessity of precision in knowing just what the proposed scale is to measure. Bedrooms, kitchen, dining room, basement, etc., are not considered. The study is restricted to the significance of the living room in the home. Such careful definition of purpose is a necessity for worth-while results. The Living Room Scale is reproduced below:

SCALE FOR RATING LIVING ROOM EQUIPMENT

DIRECTIONS TO VISITOR

1. The following list of items is for the guidance of the recorder. Not all of the features listed will be found in any one home. Entries on the schedules should, however, follow the order and numbering indicated. Weights appear after the names of the respective items. Disregard these weights in recording. Only when the list is finally checked should the individual items be multiplied by these weights

[4] Chapin, F. S., "Socio-Economic Status: Some Preliminary Results of Measurement," *Amer. Jour. Soc.*, January, 1932, p. 581.

and the sum of the weighted scores be computed, and then only after leaving the home. All information is confidential.

2. Check or underline the articles or items present. If more than one, write 2, 3, or 4, as the case may be.
3. Do not enter the *score* of any article or feature present. Complete recording before attempting to enter scores.
4. In cases where the family has no real living room, but uses the room at nights as a bedroom, or during the day as a kitchen or as a dining room, or as both, *in addition to use of room as the chief gathering place of the family, please note this fact clearly* and describe for what purposes the room is used.
5. When possible it is desirable to have a living room checked twice. This may be done in either of two ways.
   a. After an interval of two or three weeks the same visitor may recheck the room. The first schedule should be marked I, the second II.
   b. After an interval or simultaneously the room may be checked by two different visitors. One schedule should be marked A, the other B.

Scores of the same homes on two trials should be similar. If a group of homes are scored twice there should be a high correlation between the scores. Please report findings to F. Stuart Chapin, University of Minnesota.

SCHEDULE OF LIVING ROOM EQUIPMENT

I. Fixed Features

1. Floor.................... ______
   Softwood 1, hardwood 2, composition 3, stone 4.
2. Floor covering............ ______
   Composition 1, carpet 2, small rugs 3, large rug 4, oriental rug 6.
3. Wall covering............. ______
   Paper 1, kalsomine 2, plain paint 3, decorative paint 4, wooden panels 5.
4. Woodwork................ ______
   Painted 1, varnished 2, stained 3, oiled 4.
5. Door protection........... ______
   Screen 1, storm door 1.
6. Windows................. ______
   1 each window.
7. Window protection[1]....... ______
   Screen, blind, netting, storm sash, awning, shutter 1 each.
8. Window covering[1]........ ______
   Shades 1, curtains 2, drapes 3.
9. Fireplace................ ______
   Imitation 1, gas 2, wood 4, coal 4.
10. Fire utensils.............. ______
    Andirons, screen, poker, tongs, shovel, brush, hod, basket, rack. 1 each.
11. Heat.................... ______
    Stove 1, hot air 2, steam 3, hot water 4.
12. Artificial light.
    Kerosene 1, gas 2, electric 3.
13. Artificial ventilators 1......_
14. Clothes closets 1..........._
    Total Section I.........._

II. Standard Furniture

15. Table....................._
    Sewing 1, writing 1, card 1, library, end, tea, 2 each.
16. Chair....................._
    Straight, rocker, arm chair, high chair, 1 each.
17. Stool or bench............._
    High stool, footstool, piano stool, piano bench, 1 each.
18. Couch...................._
    Cot 1, sanitary couch 2, chaise longue 3, daybed 4, davenport 5, bed-davenport 6.

19. Desk .... ___
Business 1, personal-social 2.
20. Book case 1 .... ___
21. Wardrobe or movable cabinet 1 .... ___
22. Sewing cabinet 1 .... ___
23. Sewing machine .... ___
Hand power 1, foot power 2, electric 3.
24. Rack or stand 1 .... ___
25. Screen 1 .... ___
26. Chests 1 .... ___
27. Music cabinet 1 .... ___
Total Section II .... ___

III. Furnishings and Cultural Resources

28. Covers .... ___
Furniture, table, chair, couch, piano, 1 each.
29. Pillows .... ___
Couch, floor, 1 each.
30. Lamps .... ___
Floor, bridge, table, 1 each.
31 Candle holders, 1 each ..... ___
32. Clock .... ___
Mantel, grandfather, wall, alarm, 1 each.
33. Mirror, 1 each .... ___
34. Pottery, brass or metal ..... ___
Factory made 1, hand made 2 each.
35. Baskets .... ___
Factory or hand made, waste, sewing, sandwich, decorative, 1 each.
36. Statues 1 each .... ___
37. Vases 1, flowers or plants, 2 each .... ___
38. Photographs 1 each (portraits of personal interest) .. ___
39. Pictures .... ___
Note if original or reproduction. If original, oil, water color, etching, wood block, lithograph, crayon drawing, pencil drawing, pen and ink, brush drawing, photograph (when treated as a work of art), 2 each; if reproduction, photograph, half tone, color print, chromo, 1 each.
40. Books [2] ___
Poetry, fiction, history, drama, biography, philosophy, essays, literature, religion, art, science (physical, psychological, social), atlas, dictionary, encyclopedia, .20 for each volume.
41. Newspapers [3] .... ___
General, labor, local community, sectarian, 1 for each type of paper.
42. Periodicals [3] .... ___
News (current events), professional, religious, literary, science, art, children's, 1 each; fraternal, fashion, or popular story, .50 each.
43. Telephone [3] .... ___
Switchboard connection 1, two-party line 2, one-party line 3 (Note social or business mainly.)
44. Radio [3] .... ___
Crystal 1, one-tube 2, two-tube 3, three-tube 4, five-tube and up, 5.
45. Musical instruments [3] ...... ___
Piano 5, organ 1, violin 1, other hand instruments, 1 each.
46. Mechanical musical instruments [3] .... ___
Music box 1, phonograph 2, player-organ 3, player-piano 4.
47. Sheet music [3] .... ___
Opera, folk, military, ballads, classic, dance (other than jazz), children's exercises, .05 for each sheet; jazz, .01 for each sheet.
48. Phonograph records [3] ...... ___
Type of music (as above); type of instrument reproduced; voice—solo, duet, quartet, chorus; instrumental—solo, instrument (piano, violin, etc.), trio, quartet, band, orchestra, .10 for each record; jazz, .01 for each.
Total Section III .... ___

IV. Atmosphere and "Gestalt" of Room

49. Cleanliness of room and furnishings
a. Spotted or stained (−4)
b. Dusty (−2) ___
c. Spotless and dustless

[1] If checked out of season, ascertain if used in season and so record.

[2] To be recorded if in another room (except professional library of doctor, lawyer, clergyman).

[3] To be recorded if in another room.

50. Orderliness of room and furnishings
    a. Articles strewn about in disorder(−2)____
    b. Articles in place or in usable order (+2) ____
51. Condition of repair of articles and furnishings
    a. Broken, scratched, frayed, ripped, or torn (−4)____
    b. Articles or furnishings patched up (−2) ____
    c. Articles or furnishings in good repair and well kept (+2) ____
52. Record your general impression of good taste
    a. Bizarre, clashing, inharmonious or offensive (−4)____
    b. Drab, monotonous, neutral, inoffensive (−2)____
    c. Attractive in a positive way, harmonious, quiet and restful (+2) ____

Total Section IV..........____

Sums of Weighted Scores

Total Section I..............____
Section II..............____
Section III..............____
Section IV..............____
Grand Total..................____

How Is the Living Room Scale Related to Other Criteria

This scale makes it possible to measure home environment in terms of socio-economic status. The original study upon which the scale is based, defined socio-economic status as the position that an individual or a family occupies with reference to the prevailing average standards of cultural possessions, effective income, material possessions, and participation in group activity of the community. Effective income was measured by the Sydenstricker-King scale; cultural possessions, material possessions, and participation in group activity of the community were each measured by separately devised scales. The living room scale was then constructed as a simple measure which showed high correlation with the composite scores of the original four measures of socio-economic status.

VALIDITY

(1) 38 homes with Chapman-Sims scale, $r = +.69 \pm .08$
(2) 18 homes with Holley, $\rho = +.514$
(3) 29 Minnesota Children's Bureau cases with social worker's judgments, biserial $r = +.90$
(4) 75 homes in New York with 60 environmental factors (Van Alstyne, p. 59) $r = +.68 \pm .04$.

RANGES OF SCALE

Tester	Place	Number	Range	Mean	Social Class
Chapin	Minneapolis	38	20– 89	50	Middle class[4]
Taeuber	Minneapolis	46	60–359	163	Upper middle[4]
Chapin	Twin Cities	29	25–108	62	Middle class[4]
Van Alstyne	New York	75	20–200	76	Middle class[5]
Conklin	Brooklyn, N. Y.	128	44–384	111	Upper middle[5]

CORRELATION WITH OTHER FACTORS

Factor	Correlation	No. Cases	Investigator
Education of parents	$r = +.71$	120	Skalet
Occupational status (Minnesota Occupational Classification)	$r = +.74$	120	Skalet

Factor	Correlation	No. Cases	Investigator
I. Q. of child[6]	$r = +.46$	70	Skalet
Child's M.A.[7]	$r = +.59$	75	Van Alstyne
Mother's intelligence	$r = +.65$	75	Van Alstyne
Child's vocabulary[7]	$r = +.67$	75	Van Alstyne

[4] Detached and duplex houses.
[5] Flats and apartment houses.
[6] Four-year-olds.
[7] Three-year-olds.

The Scale is divided into four sections, and each item has a weight assigned to it. The sums of the weighted scores for each section, then, give the relative importance of a particular home. The grand total weighted score gives a basis of comparing one living room with another and of arranging a large number of such measures in the form of a frequency distribution to compare one community with another. The Scale has been used to measure the socio-economic status of over six hundred homes, and a project is now under way to standardize it.[5]

4. WOODWORTH-MATTHEWS PSYCHONEUROTIC INVENTORY

In 1918, Professor R. S. Woodworth developed the Psychoneurotic Inventory for the purpose of detecting psychopathic and neurotic tendencies among the soldiers of the American army. The original Inventory contained 116 questions, or statements. Dr. Ellen Matthews eliminated 46 of the statements to adapt it to use with school children. That left 70 statements, the form in which she used it for investigations among normal school children.[6] In his study of delinquent boys in four institutions in New York State, Dr. John Slawson used the Inventory to determine the psycho-neurotic status of delinquent boys as compared with that of non-delinquents. The Inventory is reproduced below:

MATTHEWS REVISION OF THE PSYCHONEUROTIC INVENTORY

1. Do you like to play by yourself better than to play with other boys?.............................. Yes No
2. Do other boys let you play with them?........... Yes No
3. Did you ever run away from home?............... Yes No

[5] Chapin, *op. cit.*, pp. 581, 586, 587.
[6] Matthews, Ellen, "A Study of Emotional Stability in Children," *Jour. Delinq.*, 1921, No. 8, pp. 1-40.
The Inventory, as given below, has been taken from Slawson, John, *The Delinquent Boy*, pp. 218-221. Boston: Richard G. Badger, 1926.

4. Did you ever want to run away from home?........ Yes No
5. Do people find fault with you much?............... Yes No
6. Do you think people like you as much as they do other people? .... Yes No
7. Does it make you uneasy to cross a bridge over water? Yes No
8. Do you mind going into a tunnel or subway?........ Yes No
9. Are you afraid of water?.......................... Yes No
10. Are you afraid during a thunder storm?............. Yes No
11. Do you feel like jumping off when you are on a high place? .... Yes No
12. Are you afraid of the dark?....................... Yes No
13. Are you often frightened in the middle of the night?.. Yes No
14. Do you have a light in your room at night?.......... Yes No
15. Do you ever cry out in your sleep?................ Yes No
16. Do you talk in your sleep?........................ Yes No
17. Do you walk in your sleep?....................... Yes No
18. Are you troubled with dreams about your play?...... Yes No
19. Do you ever have the same dream over and over?.... Yes No
20. Do you ever cry yourself to sleep?................ Yes No
21. Did you ever have the habit of picking your toes or your nose? .... Yes No
22. Did you ever have the habit of stuttering?.......... Yes No
23. Can you sit still without fidgeting?................ Yes No
24. Did you ever have the habit of twitching your head, neck or shoulders?............................ Yes No
25. Do you break and tear and spoil things more than other people? .... Yes No
26. Do you ever get so angry that you see red?......... Yes No
27. Do you stumble and fall over things more than other people? .... Yes No
28. Are you usually happy?........................... Yes No
29. Do you ever feel that nobody loves you?............ Yes No
30. Do you ever wish you had never been born?........ Yes No
31. Do you ever wish you were dead?.................. Yes No
32. Do you ever giggle over nothing at all?............ Yes No
33. Is it easy to get you cross over very small things?.... Yes No
34. Did you ever have a real fight?.................... Yes No
35. Do you like to tease people till they cry?........... Yes No
36. Can you stand pain as quietly as others do?........ Yes No
37. Do you ever feel a certain pleasure in hurting a person or an animal? .... Yes No
38. Do you feel that you are a little bit different from other people?.............................. Yes No

39. Do you seem to have a harder time to get along in school than other boys do?........................ Yes No
40. Do you ever feel that your parents are not really your own? .... Yes No
41. Do you ever have the feeling as if you were falling just before going to sleep?........................ Yes No
42. Do you ever feel as if you were smothering?........ Yes No
43. Are you usually on time?.......................... Yes No
44. Do you usually feel well and strong?............... Yes No
45. Do you usually sleep well?......................... Yes No
46. Do you feel well rested in the morning?,............ Yes No
47. Do you feel sort of tired a good deal of the time?.... Yes No
48. Do you feel bored a good deal of the time?......... Yes No
49. Do your eyes often pain you?...................... Yes No
50. Do you have many bad headaches?................ Yes No
51. Have you ever fainted away?....................... Yes No
52. Does your family treat you right?.................. Yes No
53. Do your teachers generally treat you right?........ Yes No
54. Are you ever bothered by a feeling that things are not real? .... Yes No
55. Are you ever troubled with the idea that somebody is following you?.............................. Yes No
56. Do you ever feel that someone is trying to do you harm? .... Yes No
57. Does it make you uneasy to cross a wide street or open square?.................................. Yes No
58. Does it make you uneasy to sit in a small room with the door shut?................................ Yes No
59. Do you usually know just what you want to do next? Yes No
60. Do you have a hard time making up your mind about things? .... Yes No
61. Do you have a great fear of fire?.................. Yes No
62. Do you ever have a strong desire to set fire to something? .... Yes No
63. Did you ever have a strong desire to steal things?.... Yes No
64. Do you think you have more fears than most people?.. Yes No
65. Do you make friends easily?....................... Yes No
66. Do you get tired of people easily?.................. Yes No
67. Have you any very strong superstitions?........... Yes No
68. Did you ever have a vision?....................... Yes No
69. Did you ever feel that you were very wicked?........ Yes No
70. Do you consider yourself a very moody person?...... Yes No

The strong points in favor of the Inventory are the simplicity of the questions asked, the fact that the answers are either "yes" or "no," the fact that the answers can be treated quantitatively by regarding the verbal responses as symptomatic of mental states, and the further fact that the scoring is not dependent upon the opinions of the scorer. While no claim is made that the Inventory is better than a psychiatric examination or that it should supersede such an examination, it has an advantage in that it provides a basis for quantitative comparison of the reactions of a special group, such as delinquent boys, with the reactions of an unselected group of non-delinquent children. The scores can be arranged in a frequency distribution and compared either graphically or in terms of averages and dispersions. However, the Inventory is open to all the criticisms to which any questionnaire is liable. There is no way of checking the veracity of the answers, and there is no way of knowing whether boys of all grades of intellectual ability understand the questions alike. The reliability of the results of the Inventory depends upon the veracity of the answers and the question of uniformity of understanding.[7]

Every such scoring device as the Inventory requires "standardization." The technique of standardizing a scale involves two procedures. First, the scale should be applied by different observers to the same subject or subjects. How closely do the ratings of the different observers agree? The degree of correlation between the results is a measure of the reliability, or internal consistency, of the rating scale. That is, if the coefficient of correlation is high, it indicates that different observers can apply the scale in the same way and obtain similar results. Second, other scales should be applied to the same subject or subjects. How closely do the ratings agree? The degree of correlation between results obtained on each of the rating scales and results obtained on the scale to be standardized is a measure of validity, or external consistency, of the rating scales. If the correlation is low, the question arises as to which scale is better. Obviously they do not measure the same thing, or, if they do, they do not measure it in the same way. Some of the other tests suggested by Gould may then be applied to the scale.[8]

[7] Slawson has discussed the strength and the weakness of the Inventory and concludes that his results are reasonably reliable. See *op. cit.*, pp. 221-223.

[8] See pp. 494, 495 above.

It is important at an early stage in the use of a scale to determine the form of distribution of the trait in question. Is it distributed according to the normal curve of error, or is it distributed in the form of a skewed curve? That is, a norm must be established with which to compare other sample studies. This could be accomplished by taking a random sample, or unselected group of individuals, of sufficient size and applying the scale to them. It is desirable to take several random samples as a check on the validity of the guess that any particular sample is random. If the results in each case are similar, it may be assumed that the scale has been applied to similar samples and that the form of the distribution of the trait is a satisfactory norm. This is on the assumption, of course, that the scale has been standardized for reliability and validity.

Attention should be called to the fact that each question in the Inventory is given equal weight. This raises a different problem regarding rating scales: that of weighting the questions or statements. The justification for giving equal weight to all statements may be questioned. For example, questions 9 and 10 in the Inventory are similar, but they involve stimuli which are different qualitatively and quantitatively: "Are you afraid of water?" and "Are you afraid during a thunder storm?" Do they equally reflect the mental stability of the individual? How could such a question be answered satisfactorily? Should the first be given a weight of two and the second a weight of one, or vice versa? Of course, the assumption regarding the differential importance of the questions is that with a large number of questions, many of which are similar, the necessity of a weighting scheme is eliminated. But that is an open question which the maker of rating scales should always take into account.

5. MEASUREMENT OF ATTITUDE TOWARD THE CHURCH

The scale worked out by Professors L. L. Thurstone and E. J. Chave for measuring attitudes toward the church provides a good example of the method of constructing rating scales for attitudes and of methods of standardization. In planning this scale the first problem was to determine what opinions about the church actually exist. "Several groups of people and many individuals," the authors explain, "were asked to write out their opinions about the church, and current literature was searched for suitable brief statements that might serve the purposes of the scale. By editing such

material a list of 130 statements was prepared, expressive of attitudes covering as far as possible all gradations from one end of the scale to the other."[9] Careful attention was given to selecting a list of opinions ranging all the way from complete confidence to complete antagonism. In the middle of the range would be found more or less neutral statements of opinion. Attention to the neutral opinions was of fundamental importance to prevent the scale from breaking into two parts and the scores being distributed in a U-shaped curve instead of in the form of a normal distribution.

Certain practical criteria were applied to the first editing of the work. The most important were as follows: "(1) The statements should be as brief as possible so as not to fatigue the subjects who are asked to read the whole list. (2) The statements should be such that they can be indorsed or rejected in accordance with their agreement or disagreement with the attitude of the reader. Some statements in a random sample will be so phrased that the reader can express no definite indorsement or rejection of them. (3) Every statement should be such that acceptance or rejection of the statement does indicate something regarding the reader's attitude about the issue in question. If, for example, the statement is made that war is an incentive to inventive genius, the acceptance or rejection of it really does not say anything regarding the reader's pacifistic or militaristic tendencies. He may regard the statement as an unquestioned fact and simply indorse it as a fact, in which case his answer has not revealed anything concerning his own attitude on the issue in question. However, only the conspicuous examples of this effect should be eliminated by inspection, because an objective criterion is available for detecting such statements so that their elimination from the scale will be automatic. Personal judgment should be minimized as far as possible in this type of work. (4) Double-barreled statements should be avoided except possibly as examples of neutrality when better neutral statements do not seem to be readily available. Double-barreled statements tend to have a high ambiguity. (5) One must insure that at least a fair majority of the statements really belong on the attitude variable that is to be measured. If a small number of irrelevant statements should be either intentionally or unintentionally left in the series, they will be automatically eliminated by an objective criterion, but the criterion will not be successful unless the ma-

[9] *Op. cit.*, p. 22.

jority of the statements are clearly a part of the stipulated variable."[10]

A list of 130 was taken from the statements obtained from individuals and from literature. In order to arrive at an approximate gradation of the statements ranging from highest appreciation to highest depreciation of the church, 341 individuals were asked to arrange the statements in eleven groups, beginning with highest appreciation and ending with highest depreciation. The 130 statements were mimeographed on small slips of paper, and each subject was given 11 master-slips lettered A to K. F fell in the middle, and to this master-slip was to be assigned all the statements regarded as neutral. Only the first, middle and last piles were given descriptions; within this range the subjects were to classify the opinions. The authors worked out scale values for each statement from this sorting. A few of the statements are given to illustrate the types used:[11]

1. I have seen no value in the church.
2. I believe the modern church has plenty of satisfying interests for young people.
3. I do not hear discussions in the church that are scientific or practical and so I do not care to go.
4. I believe that membership in a good church increases one's self-respect and usefulness.
5. I believe a few churches are trying to keep up to date in their thinking and methods of work, but most are far behind the times.
6. I regard the church as an ethical society promoting the best way of living for both an individual and for society.

It will be noted that Thurstone and Chave used statements which were to be marked "yes" or "no," while Woodworth-Matthews used questions. There may be a question as to which method is better, but extensive experimentation would be required to decide this. Furthermore, the two tests are seeking different things. Woodworth and Matthews are asking for a report of experience as a matter of fact, while Thurstone and Chave are asking for an expression of opinion. Thurstone and Chave regard opinions as symbolic of attitudes, and their study of attitudes is based upon the theory that an attitude is correctly represented by verbal opin-

[10] *Op. cit.*, pp. 22, 23.
[11] *Op. cit.*, Chap. II.

ions. This assumption might also be questioned, and the authors recognize that fact.

A final list of 45 statements was selected from the 130 opinions, after the criteria of ambiguity and irrelevance had been applied and after consideration of the scale values and careful inspection of the statements themselves. From this final study an "experimental attitude scale" was developed. The authors summarize their judgment regarding the scale thus: "The essential characteristic of the present measurement method is the scale of evenly graduated opinions so arranged that equal steps or intervals on the scale *seem* to most people to represent equally noticeable shifts in attitude."[12] That is, a means has been devised for treating attitudes as continuous variables. This is an important step in the quantitative treatment of facts traditionally regarded as qualitative and subjective. It shows that any dogmatic skepticism about measurements in psychology and the social sciences is of doubtful validity and, as experimentation proceeds, may be proved largely unwarranted.

6. EXERCISES

1. Devise the following types of rating scales with the individual items appropriately weighted:
   (a) A scale for rating student room equipment.
   (b) A scale for rating student attitudes toward military training in colleges.
2. Obtain from the University of Minnesota a supply of Chapin's Scale for Rating Living Room Equipment and make a survey of 100 or more living rooms in your college town. If each student does a certain number of these, the field work will not be laborious. Then the data on all schedules can be combined for analysis and comparison of homes. Compare your results with Chapin's.
3. Obtain from the University of Chicago Press a supply of Thurstone and Chave's scale for measuring attitudes toward the church and get them filled out by 100 or more students. If each student in the class takes his pro rata of the forms, the time required for obtaining the original data will not be great. Returns may be pooled for analysis by each student. Compare your results with Thurstone and Chave's.

[12] *Op. cit.*, p. 82.

7. REFERENCES

Chapin, F. S., "A Quantitative Scale for Rating the Home and Social Environment of Middle Class Families in an Urban Community: A First Approximation to the Measurement of Socio-Economic Status," *Jour. Educ. Psych.*, No. 2, pp. 99-111.

———"Socio-Economic Status: Some Preliminary Results of Measurement," *Amer. Jour. Soc.*, Vol. XXXVII, No. 4, pp. 581-587.

———"The Meaning of Measurement in Sociology," *Pub. of the Amer. Soc. Soc.*, Vol. XXIV, pp. 83-94.

Hartshorne, Hugh, and May, Mark A., *Studies in Deceit*, Chaps. III, IV, VIII, IX.

Lundberg, George A., *Social Research*, Chaps. IX and X.

McCormick, Mary J., *The Measurement of Home Conditions*, a pamphlet published by the National Catholic School of Social Service, Washington.

Slawson, John, *The Delinquent Boy*, Chap. IV.

Thurstone, L. L., and Chave, E. J., *The Measurement of Attitude.*

Watson, G. B., *The Measurement of Fair-Mindedness.*

APPENDICES

APPENDIX A

TABLE CXXI [1]

Ordinates of the normal probability curve expressed as fractional parts of the mean ordinate y_o. Each ordinate is erected at a given distance from the mean. The height of the ordinate erected at the mean can be computed from,

$$y_o = \frac{N}{\sigma\sqrt{2\pi}} = \frac{N}{2.5066\ \sigma}$$

The corresponding height of any other ordinate can be read from the table by assigning the distance that the ordinate is from the mean (x). Distances on x are measured as fractional parts of σ. Thus the height of an ordinate at a distance from the mean of $.7\sigma$ will be $.78270\ y_o$; the height of an ordinate at $2.15\ \sigma$ from the mean will be $.09914\ y_o$, etc.

x/σ	0	1	2	3	4	5	6	7	8	9
0.0	100000	99995	99980	99955	99920	99875	99820	99755	99685	99596
0.1	99501	99396	99283	99158	99025	98881	98728	98565	98393	98211
0.2	98020	97819	97609	97390	97161	96923	96676	96420	96156	95882
0.3	95600	95309	95010	94702	94387	94055	93723	93382	93024	92677
0.4	92312	91399	91558	91169	90774	90871	89961	89543	89119	88688
0.5	88250	87805	87353	86896	86432	85962	85488	85006	84519	84060
0.6	83527	83023	82514	82010	81481	80957	80429	79896	79359	78817
0.7	78270	77721	77167	76610	76048	75484	74916	74342	73769	73193
0.8	72615	72033	71448	70861	70272	69681	69087	68493	67896	67298
0.9	66689	66097	65494	64891	64287	63683	63077	62472	61865	61259
1.0	60653	60047	59440	58834	58228	57623	57017	56414	55810	55209
1.1	54607	54007	53409	52812	52214	51620	51027	50437	49848	49260
1.2	48675	48092	-47511	46933	46357	45783	45212	44644	44078	43516
1.3	42956	42399	41845	41294	40747	40202	39661	39123	38569	38058
1.4	37531	37007	36487	35971	35459	34950	34445	33944	33447	32954
1.5	32465	31980	31500	31023	30550	30082	29618	29158	28702	28251
1.6	27804	27361	26923	26489	26059	25634	25213	24797	24385	23978
1.7	23575	23176	22782	22392	22008	21627	21251	20879	20511	20148
1.8	19790	19436	19086	18741	18400	18064	17732	17404	17081	16762
1.9	16448	16137	15831	15530	15232	14939	14650	14364	14083	13806
2.0	13534	13265	13000	12740	12483	12230	11981	11737	11496	11259
2.1	11025	10795	10570	10347	10129	09914	09702	09495	09290	09090
2.2	08892	08698	08507	08320	08136	07956	07778	07604	07433	07265
2.3	07100	06939	06780	06624	06471	06321	06174	06029	05888	05750
2.4	05614	05481	05350	05222	05096	04973	04852	04734	04618	04505
2.5	04394	04285	04179	04074	03972	03873	03775	03680	03586	03494
2.6	03405	03317	03232	03148	03066	02986	02908	02831	02757	02684
2.7	02612	02542	02474	02408	02343	02280	02218	02157	02098	02040
2.8	01984	01929	01876	01823	01772	01723	01674	01627	01581	01536
2.9	01492	01449	01408	01367	01328	01288	01252	01215	01179	01145
3.0	01111	00819	00598	00432	00309	00219	00153	00106	00073	00050
4.0	00034	00022	00015	00010	00006	00004	00003	00002	00001	00001
5.0	00000									

[1] Rugg, H. O., *Statistical Methods Applied to Education*, p. 388. Boston: Houghton Mifflin Co., 1917. Reprinted by permission of the publishers.

TABLE CXXII[1]

Fractional parts of the total area (10,000) under the normal probability curve, corresponding to distances on the baseline between the mean and successive points of division laid off from the mean. Distances are measured in units of the standard deviation, σ. To illustrate, the table is read as follows: between the mean ordinate, y_o, and any ordinate erected at a distance from it of, say, $.8\sigma$ $\left(i.e., \frac{x}{\sigma} = .8\right)$, is included 28.81 per cent of the entire area.

x/σ	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
0.0	0000	0040	0080	0120	0159	0199	0239	0279	0319	0359
0.1	0398	0438	0478	0517	0557	0596	0636	0675	0714	0753
0.2	0793	0832	0871	0910	0948	0987	1026	1064	1103	1141
0.3	1179	1217	1255	1293	1331	1368	1406	1443	1480	1517
0.4	1554	1591	1628	1664	1700	1736	1772	1808	1844	1879
0.5	1915	1950	1985	2019	2054	2088	2123	2157	2190	2224
0.6	2257	2291	2324	2357	2389	2422	2454	2486	2518	2549
0.7	2580	2612	2642	2673	2704	2734	2764	2794	2823	2852
0.8	2881	2910	2939	2967	2995	3023	3051	3078	3106	3133
0.9	3159	3186	3212	3238	3264	3289	3315	3340	3365	3389
1.0	3413	3438	3461	3485	3508	3531	3554	3577	3599	3621
1.1	3643	3665	3686	3718	3729	3749	3770	3790	3810	3830
1.2	3849	3869	3888	3907	3925	3944	3962	3980	3997	4015
1.3	4032	4049	4066	4083	4099	4115	4131	4147	4162	4177
1.4	4192	4207	4222	4236	4251	4265	4279	4292	4306	4319
1.5	4332	4345	4357	4370	4382	4394	4406	4418	4430	4441
1.6	4452	4463	4474	4485	4495	4505	4515	4525	4535	4545
1.7	4554	4564	4573	4582	4591	4599	4608	4616	4625	4633
1.8	4641	4649	4656	4664	4671	4678	4686	4693	4699	4706
1.9	4713	4719	4726	4832	4738	4744	4750	4758	4762	4767
2.0	4773	4778	4783	4788	4793	4798	4803	4808	4812	4817
2.1	4821	4826	4830	4834	4838	4842	4846	4850	4854	4857
2.2	4861	4865	4868	4871	4875	4878	4881	4884	4887	4890
2.3	4893	4896	4898	4901	4904	4906	4909	4911	4913	4916
2.4	4918	4920	4922	4925	4927	4929	4931	4932	4934	4936
2.5	4938	4940	4941	4943	4945	4946	4948	4949	4951	4952
2.6	4953	4955	4956	4957	4959	4960	4961	4962	4963	4964
2.7	4965	4966	4967	4968	4969	4970	4971	4972	4973	4974
2.8	4974	4975	4976	4977	4977	4978	4879	4980	4980	4981
2.9	4981	4982	4983	4984	4984	4984	4985	4985	4986	4986

[1] Rugg, H. O., *op. cit.*, p. 389.

TABLE CXXIII

Tables of the Chi-Function for the Pearson Chi Test[1]

χ^2	$n = 3$	$n = 4$	$n = 5$	$n = 6$
1	.60653 06597	.80125 195(69)	.90979 598(96)	.96256 577(32)
2	.36787 94412	.57240 670(44)	.73575 888(23)	.84914 503(60)
3	.22313 01601	.39162 517(63)	.55782 540(04)	.69998 583(59)
4	.13533 52832	.26146 412(99)	.40600 584(97)	.54941 595(12)
5	.08208 49986	.17179 714(43)	.28729 749(52)	.41588 018(72)
6	.04978 70684	.11161 022(51)	.19914 827(35)	.30621 891(86)
7	.03019 73834	.07189 777(25)	.13588 822(54)	.22064 030(80)
8	.01831 56389	.04601 170(57)	.09157 819(44)	.15623 562(76)
9	.01110 89965	.02929 088(65)	.06109 948(10)	.10906 415(79)
10	.00673 79470	.01856 612(57)	.04042 768(20)	.07523 523(64)
11	.00408 67714	.01172 587(55)	.02656 401(44)	.05137 998(34)
12	.00247 87522	.00738 316(05)	.01735 126(52)	.03478 778(05)
13	.00150 34392	.00463 660(55)	.01127 579(39)	.02337 876(81)
14	.00091 18820	.00290 515(28)	.00729 505(57)	.01560 941(61)
15	.00055 30844	.00181 664(90)	.00470 121(71)	.01036 233(79)
16	.00033 54626	.00113 398(42)	.00301 916(37)	.00684 407(35)
17	.00020 34684	.00070 674(24)	.00193 294(95)	.00449 979(70)
18	.00012 34098	.00043 984(97)	.00123 409(80)	.00294 640(46)
19	.00007 48518	.00027 339(89)	.00078 594(42)	.00192 213(68)
20	.00004 53999	.00016 974(16)	.00049 939(92)	.00124 972(97)
21	.00002 75364	.00010 527(62)	.00031 666(92)	.00081 005(96)
22	.00001 67017	.00006 523(11)	.00020 042(04)	.00052 359(83)
23	.00001 01301	.00004 038(30)	.00012 662(62)	.00033 756(61)
24	.00000 61442	.00002 498(00)	.00007 987(48)	.00021 711(29)
25	.00000 37267	.00001 544(05)	.00005 030(98)	.00013 933(73)
26	.00000 22603	.00000 953(74)	.00003 164(46)	.00008 923(60)
27	.00000 13710	.00000 600(96)	.00001 987(89)	.00005 716(47)
28	.00000 08315	.00000 361(89)	.00001 247(29)	.00003 638(57)
29	.00000 05043	.00000 223(94)	.00000 781(74)	.00002 318(76)
30	.00000 03059	.00000 137(09)	.00000 489(44)	.00001 473(95)
40	.00000 00021	.00000 001(07)	.00000 004(12)	.00000 014(93)
50	.00000 00000	.00000 000(00)	.00000 000(03)	.00000 000(13)
60	.00000 00000	.00000 000(00)	.00000 000(00)	.00000 000(00)
70	.00000 00000	.00000 000(00)	.00000 000(00)	.00000 000(00)

[1] Computed by Miss Anna M. Lescisin, Indiana University, to 10 decimal places. The last two places in parentheses indicate some lack of confidence in these figures. The following errors are to be noted:

Pearson's Value		Our Value	
χ^2	$n = 12$	χ^2	12
7	.799073	7	.799083
12	.362642	12	.363643

χ^2	$n = 7$	$n = 8$	$n = 9$	$n = 10$
1	.98561 232(20)	.99482 853(65)	.99824 837(74)	.99943 750(26)
2	.91969 860(29)	.95984 036(87)	.98101 184(31)	.99146 760(65)
3	.80884 683(05)	.83500 223(17)	.93435 754(56)	.96429 497(27)
4	.67667 641(62)	.77977 740(84)	.85712 346(05)	.91141 252(67)
5	.54381 311(59)	.65996 323(00)	.75757 613(31)	.83430 826(07)
6	.42319 008(11)	.53974 935(08)	.64723 188(88)	.73991 829(27)
7	.32084 719(89)	.42887 985(77)	.53663 266(80)	.63711 940(74)
8	.23810 330(56)	.33259 390(26)	.43347 012(03)	.53414 621(68)
9	.17357 807(09)	.25265 604(65)	.34229 595(58)	.43727 418(87)
10	.12465 201(95)	.18857 345(78)	.26502 591(53)	.35048 520(26)
11	.08837 643(24)	.13861 902(08)	.20169 919(87)	.27570 893(67)
12	.06196 880(44)	.10055 886(85)	.15120 388(28)	.21330 930(51)
13	.04303 594(69)	.07210 839(10)	.11184 961(16)	.16260 626(22)
14	.02963 616(39)	.05118 135(34)	.08176 541(63)	.12232 522(80)
15	.02025 671(51)	.03599 940(48)	.05914 545(98)	.09093 597(66)
16	.01375 396(77)	.02511 635(89)	.04238 011(41)	.06688 158(26)
17	.00928 324(43)	.01739 618(25)	.03010 907(97)	.04871 597(63)
18	.00623 219(51)	.01197 000(23)	.02122 648(63)	.03517 353(94)
19	.00416 363(30)	.00818 734(10)	.01485 964(77)	.02519 289(50)
20	.00276 939(57)	.00556 968(23)	.01033 605(07)	.01791 240(37)
21	.00183 461(59)	.00377 015(01)	.00714 742(96)	.01265 042(13)
22	.00121 087(33)	.00254 041(40)	.00491 586(73)	.00887 897(75)
23	.00079 647(86)	.00170 458(70)	.00336 424(63)	.00619 629(64)
24	.00052 225(81)	.00113 935(12)	.00229 179(12)	.00430 131(09)
25	.00034 145(46)	.00075 880(38)	.00155 455(79)	.00297 118(41)
26	.00022 264(24)	.00050 366(86)	.00105 029(97)	.00204 298(97)
27	.00014 480(76)	.00033 340(23)	.00070 698(65)	.00139 889(00)
28	.00009 396(27)	.00021 987(94)	.00047 424(85)	.00095 385(41)
29	.00006 083(69)	.00014 468(69)	.00031 709(81)	.00064 804(12)
30	.00003 930(84)	.00009 495(06)	.00021 137(85)	.00043 871(26)
40	.00000 045(34)	.00000 125(87)	.00000 320(16)	.00000 759(84)
50	.00000 000(47)	.00000 001(44)	.00000 004(09)	.00000 010(77)
60	.00000 000(00)	.00000 000(02)	.00000 000(05)	.00000 000(13)
70	.00000 000(00)	.00000 000(00)	.00000 000(00)	.00000 000(00)

χ^2	$n = 11$	$n = 12$	$n = 13$	$n = 14$
1	.99982 788(44)	.99994 961(00)	.99998 583(51)	.99999 616(52)
2	.99634 015(31)	.99849 588(16)	.99940 581(51)	.99977 374(98)
3	.98142 406(38)	.99072 588(63)	.99554 401(93)	.99793 431(73)
4	.94734 698(27)	.96991 702(37)	.98343 639(15)	.99119 138(63)
5	.89117 801(89)	.93116 661(10)	.95797 896(18)	.97519 313(39)
6	.81526 324(46)	.87336 425(39)	.91608 205(80)	.94615 296(01)
7	.72544 495(35)	.79908 350(16)	.85761 355(34)	.90215 156(16)
8	.62883 693(51)	.71330 382(93)	.78513 038(69)	.84360 027(48)
9	.53210 357(63)	.62189 233(10)	.70293 043(47)	.77294 553(83)
10	.44049 328(51)	.53038 714(13)	.61596 065(48)	.69393 435(82)
11	.35751 800(24)	.44326 327(82)	.52891 868(64)	.61081 761(97)
12	.28505 650(03)	.36364 322(05)	.44567 964(13)	.52764 385(54)
13	.22367 181(68)	.29332 540(93)	.36904 068(36)	.44781 167(41)
14	.17299 160(79)	.23299 347(74)	.30070 827(62)	.37384 397(66)
15	.13206 185(63)	.18249 692(96)	.24143 645(10)	.30735 277(37)
16	.09963 240(69)	.14113 086(91)	.19123 607(53)	.24912 983(01)
17	.07436 397(98)	.10787 558(68)	.14959 731(00)	.19930 407(58)
18	.05496 364(15)	.08158 061(36)	.11569 052(09)	.15751 946(23)
19	.04026 268(23)	.06109 350(92)	.08852 844(83)	.12310 366(09)
20	.02925 268(81)	.04534 067(37)	.06708 596(29)	.09521 025(54)
21	.02109 356(56)	.03337 105(44)	.05038 045(10)	.07292 862(65)
22	.01510 460(07)	.02437 324(38)	.03751 981(41)	.05536 177(64)
23	.01074 657(84)	.01767 510(94)	.02772 594(22)	.04167 626(37)
24	.00760 039(07)	.01273 320(34)	.02034 102(96)	.03113 005(98)
25	.00534 550(55)	.00911 668(47)	.01482 287(47)	.02308 373(18)
26	.00374 018(59)	.00648 991(72)	.01073 388(99)	.01700 083(68)
27	.00260 434(03)	.00459 532(06)	.00772 719(57)	.01244 118(45)
28	.00180 524(88)	.00323 733(11)	.00553 204(96)	.00904 981(79)
29	.00124 604(48)	.00226 996(07)	.00393 999(04)	.00654 593(03)
30	.00085 664(12)	.00158 458(60)	.00279 242(92)	.00470 969(53)
40	.00001 694(26)	.00003 577(50)	.00007 190(68)	.00013 823(54)
50	.00000 026(69)	.00000 062(59)	.00000 139(71)	.00000 298(14)
60	.00000 000(36)	.00000 000(93)	.00000 002(26)	.00000 005(25)
70	.00000 000(00)	.00000 000(01)	.00000 000(03)	.00000 000(08)

χ^2	$n = 15$	$n = 16$	$n = 17$	$n = 18$
1	.99999 899(76)	.99999 974(64)	.99999 993(78)	.99999 998(51)
2	.99991 675(88)	.99997 034(49)	.99998 975(08)	.99999 655(76)
3	.99907 400(81)	.99959 780(14)	.99983 043(43)	.99993 049(82)
4	.99546 619(45)	.99773 734(40)	.99890 328(10)	.99948 293(27)
5	.98581 268(80)	.99212 641(19)	.99575 330(45)	.99777 083(79)
6	.96649 146(48)	.97974 774(76)	.98809 549(63)	.99318 566(26)
7	.93471 190(33)	.95764 974(76)	.97326 107(83)	.98354 890(12)
8	.88932 602(14)	.92378 270(28)	.94886 638(40)	.96654 676(94)
9	.83105 057(86)	.87751 745(11)	.91341 352(82)	.94026 179(87)
10	.76218 346(30)	.81973 990(96)	.86662 832(59)	.90361 027(73)
11	.68603 598(02)	.75259 437(02)	.80948 528(25)	.85656 398(72)
12	.60630 278(23)	.67902 905(67)	.74397 976(03)	.80013 721(78)
13	.52652 362(26)	.60229 793(88)	.67275 778(02)	.73618 603(49)
14	.44971 105(59)	.52552 912(95)	.59871 383(57)	.66710 193(89)
15	.37815 469(44)	.45141 720(81)	.52463 852(65)	.59548 164(24)
16	.31337 429(98)	.38205 162(82)	.45296 084(21)	.52383 487(84)
17	.25617 786(12)	.31886 440(74)	.38559 710(17)	.45436 611(65)
18	.20678 083(99)	.26266 556(05)	.32389 696(44)	.38884 087(72)
19	.16494 924(43)	.21373 388(26)	.26866 318(18)	.32853 216(35)
20	.13014 142(10)	.17193 268(88)	.22022 064(68)	.27422 926(67)
21	.10163 250(05)	.13682 931(99)	.17851 057(49)	.22629 029(06)
22	.07861 437(21)	.10780 390(86)	.14319 153(47)	.18471 903(57)
23	.06026 972(28)	.08413 984(45)	.11373 450(53)	.14925 066(84)
24	.04582 230(72)	.06509 348(69)	.08950 449(75)	.11943 497(03)
25	.03456 739(39)	.04994 343(75)	.06982 546(38)	.09470 961(38)
26	.02588 691(53)	.03802 267(61)	.05402 824(82)	.07446 053(08)
27	.01925 362(03)	.02873 644(02)	.04148 315(34)	.05806 790(06)
28	.01422 795(80)	.02156 902(04)	.03161 977(49)	.04493 819(83)
29	.01045 035(87)	.01608 463(15)	.02393 612(18)	.03452 612(06)
30	.00763 189(92)	.01192 148(60)	.01800 219(20)	.02634 506(73)
40	.00025 512(04)	.00045 339(40)	.00077 858(80)	.00129 409(44)
50	.00000 610(63)	.00001 204(12)	.00002 292(48)	.00004 224(03)
60	.00000 018(95)	.00000 025(22)	.00000 059(55)	.00000 105(09)
70	.00000 000(19)	.00000 000(37)	.00000 001(00)	.00000 002(16)

χ^2	$n = 19$	$n = 20$	$n = 21$	$n = 22$
1	.99999 999(66)	.99999 999(92)	.99999 999(98)	.99999 999(99)
2	.99999 887(48)	.99999 964(15)	.99999 988(85)	.99999 996(61)
3	.99997 226(42)	.99998 920(94)	.99999 590(25)	.99999 847(96)
4	.99976 255(27)	.99989 365(95)	.99995 350(19)	.99998 012(83)
5	.99885 974(71)	.99943 096(32)	.99972 264(79)	.99986 783(83)
6	.99619 700(81)	.99792 845(61)	.99889 751(20)	.99942 618(03)
7	.99012 634(23)	.99421 325(85)	.99668 505(61)	.99814 223(22)
8	.97863 656(53)	.98667 098(89)	.99186 775(69)	.99514 434(45)
9	.95974 268(74)	.97347 939(45)	.98290 726(70)	.98921 404(51)
10	.93190 636(53)	.95294 578(77)	.96817 194(28)	.97891 184(58)
11	.89435 667(78)	.92383 844(53)	.94622 253(05)	.96278 681(57)
12	.84723 749(38)	.88562 533(15)	.91607 598(28)	.93961 782(44)
13	.79157 303(33)	.83857 104(69)	.87738 404(94)	.90862 395(00)
14	.72909 126(79)	.78369 131(12)	.83049 593(74)	.86959 927(03)
15	.66196 711(92)	.72259 731(97)	.77640 761(31)	.82295 180(17)
16	.59254 738(44)	.65727 793(65)	.71662 431(09)	.76965 103(81)
17	.52310 504(49)	.58986 782(45)	.65297 365(78)	.71110 620(38)
18	.45565 260(45)	.52243 827(24)	.58740 824(45)	.64900 422(58)
19	.39182 348(26)	.45683 612(43)	.52182 602(24)	.58514 008(51)
20	.33281 967(91)	.39457 818(17)	.45792 971(48)	.52126 125(02)
21	.27941 304(74)	.33680 090(00)	.39713 259(87)	.45894 420(52)
22	.23198 513(32)	.28425 625(90)	.34051 068(25)	.39950 988(60)
23	.19059 013(01)	.23734 178(30)	.28879 453(95)	.34397 839(55)
24	.15502 778(29)	.19615 235(87)	.24239 216(34)	.29305 853(34)
25	.12491 619(79)	.16054 222(60)	.20143 110(65)	.24716 408(41)
26	.09975 791(41)	.13018 901(46)	.16581 187(60)	.20644 904(49)
27	.07899 549(06)	.10465 316(12)	.13526 399(63)	.17085 326(84)
28	.06205 545(45)	.08342 860(90)	.10939 984(50)	.14015 131(95)
29	.04837 906(72)	.06598 513(15)	.08775 938(83)	.11400 151(65)
30	.03744 649(10)	.05179 844(62)	.06985 365(61)	.09198 799(17)
40	.00208 725(70)	.00327 221(30)	.00499 541(03)	.00743 667(32)
50	.00007 548(26)	.00013 106(12)	.00022 147(66)	.00036 480(05)
60	.00000 211(82)	.00000 386(98)	.00000 719(39)	.00001 277(17)
70	.00000 004(52)	.00000 009(19)	.00000 018(21)	.00000 035(14)

χ^2	$n = 23$	$n = 24$	$n = 25$	$n = 26$
1	.99999 999(99)	.99999 999(99)	.99999 999(99)	.99999 999(
2	.99999 998(99)	.99999 999(70)	.99999 999(91)	.99999 999(
3	.99999 944(83)	.99999 980(39)	.99999 993(18)	.99999 997(
4	.99999 169(18)	.99999 659(85)	.99999 863(54)	.99999 946(
5	.99993 837(31)	.99997 185(62)	.99998 740(15)	.99999 446(
6	.99970 766(32)	.99985 410(16)	.99992 861(35)	.99996 573(
7	.99898 060(60)	.99945 189(02)	.99971 100(82)	.99985 048(1
8	.99716 023(36)	.99837 228(95)	.99908 477(06)	.99949 505(
9	.99333 132(78)	.99595 746(68)	.99759 571(63)	.99859 619(
10	.98630 473(15)	.99127 663(54)	.99454 690(82)	.99665 263(
11	.97474 874(95)	.98318 834(31)	.98901 185(90)	.99294 559(
12	.95737 907(62)	.97047 067(75)	.97990 803(63)	.98656 781(
13	.93316 120(99)	.95199 003(28)	.96612 044(11)	.97650 129(
14	.90147 920(61)	.92687 124(27)	.94665 037(70)	.96173 244(
15	.86223 798(36)	.89463 357(45)	.92075 869(07)	.94138 255(
16	.81588 585(21)	.85526 863(92)	.88807 606(39)	.91482 870(
17	.76336 197(88)	.80925 155(83)	.84866 204(50)	.88179 377(
18	.70598 832(06)	.75748 932(86)	.80300 838(29)	.84239 071(
19	.64532 843(52)	.70122 462(06)	.75198 960(99)	.79712 054(
20	.58303 975(06)	.64191 179(15)	.69677 614(68)	.74682 530(
21	.52073 812(75)	.58108 751(03)	.63872 522(33)	.69260 965(
22	.45988 878(67)	.52025 178(10)	.57926 689(09)	.63574 402(
23	.40172 961(04)	.46077 087(57)	.51979 809(34)	.57756 335(
24	.34722 942(00)	.40380 844(65)	.46159 733(63)	.51937 357(
25	.29707 473(13)	.35028 534(37)	.40576 068(10)	.46237 366(
26	.25168 202(65)	.30086 622(54)	.35316 493(16)	.40759 869(
27	.21122 647(90)	.25596 769(19)	.30445 316(24)	.35588 462(3
28	.17568 199(16)	.21578 160(01)	.26004 108(74)	.30785 324(
29	.14486 085(38)	.18030 985(77)	.22013 096(75)	.26391 602(7
30	.11846 440(38)	.14940 162(81)	.18475 178(70)	.22428 897(
40	.01081 171(68)	.01536 897(83)	.02138 681(95)	.02916 429(1
50	.00058 646(16)	.00092 132(26)	.00141 597(28)	.00213 115(3
60	.00002 242(10)	.00003 820(56)	.00006 394(92)	.00010 455(4
70	.00000 066(14)	.00000 121(61)	.00000 218(65)	.00000 384(7

χ^2	$n = 27$	$n = 28$	$n = 29$	$n = 30$
1	.99999 999(99)	.99999 999(99)	.99999 999(99)	.99999 999(99)
2	.99999 999(99)	.99999 999(99)	.99999 999(99)	.99999 999(99)
3	.99999 999(22)	.99999 999(74)	.99999 999(92)	.99999 999(97)
4	.99999 979(27)	.99999 992(12)	.99999 997(07)	.99999 998(91)
5	.99999 771(58)	.99999 899(13)	.99999 968(01)	.99999 982(88)
6	.99998 385(11)	.99999 252(42)	.99999 659(82)	.99999 847(85)
7	.99992 404(22)	.99996 208(73)	.99998 139(75)	.99999 102(21)
8	.99972 628(29)	.99985 433(73)	.99992 367(13)	.99996 079(19)
9	.99919 486(20)	.99954 613(99)	.99974 841(25)	.99986 278(76)
10	.99798 114(85)	.99880 302(90)	.99930 201(01)	.99959 947(28)
11	.99554 911(75)	.99723 878(63)	.99831 488(07)	.99898 786(41)
12	.99117 251(63)	.99429 444(57)	.99637 150(71)	.99772 850(24)
13	.98397 335(80)	.98924 715(43)	.99289 981(64)	.99538 404(86)
14	.97300 022(67)	.98125 471(54)	.98718 860(74)	.99137 737(52)
15	.95733 413(26)	.96943 194(61)	.97843 534(91)	.98501 494(02)
16	.93620 287(18)	.95294 715(46)	.96581 936(89)	.97553 586(27)
17	.90908 299(53)	.93112 248(54)	.94858 895(54)	.96218 130(19)
18	.87577 342(96)	.90351 971(04)	.92614 923(12)	.94427 237(51)
19	.83642 970(66)	.87000 144(09)	.89813 593(12)	.92128 799(99)
20	.79155 647(69)	.83075 611(69)	.86446 442(32)	.89292 708(80)
21	.74196 393(21)	.78628 826(28)	.82534 904(31)	.85914 939(95)
22	.68869 681(98)	.73737 720(58)	.78129 137(50)	.82018 942(45)
23	.63294 705(64)	.68501 243(77)	.73304 036(98)	.77654 313(69)
24	.57596 525(26)	.63031 609(48)	.68153 563(69)	.72893 166(96)
25	.51897 521(19)	.57446 199(50)	.62783 533(79)	.67824 748(16)
26	.46310 474(55)	.51860 045(36)	.57304 455(93)	.62549 104(05)
27	.40933 318(11)	.46379 491(03)	.51824 704(67)	.57170 519(67)
28	.35846 003(25)	.41097 348(97)	.46444 966(56)	.51791 300(14)
29	.31108 235(48)	.36089 918(32)	.41252 813(30)	.46506 627(69)
30	.26761 101(60)	.31415 380(21)	.36321 781(87)	.41400 360(46)
40	.03901 199(08)	.05123 679(26)	.06612 763(88)	.08393 679(44)
50	.00314 412(10)	.00455 081(48)	.00646 748(31)	.00903 166(94)
60	.00016 776(98)	.00026 379(32)	.00040 735(59)	.00061 765(60)
70	.00000 663(45)	.00001 121(69)	.00001 861(00)	.00003 032(18)

APPENDIX B

TABLE CXXIV

TABLE OF SQUARES, SQUARE ROOTS, AND RECIPROCALS, 1 TO

No.	Square	Square Root	Reciprocal	No.	Square	Square Root	Reciprocal
1	1	1.0000000	1.000000000	51	26 01	7.1414284	.019607843
2	4	1.4142136	0.500000000	52	27 04	7.2111026	.019230769
3	9	1.7320508	.333333333	53	28 09	7.2801099	.018867925
4	16	2.0000000	.250000000	54	29 16	7.3484692	.018518519
5	25	2.2360680	.200000000	55	30 25	7.4161985	.018181818
6	36	2.4494897	.166666667	56	31 36	7.4833148	.017857143
7	49	2.6457513	.142857143	57	32 49	7.5498344	.017543860
8	64	2.8284271	.125000000	58	33 64	7.6157731	.017241379
9	81	3.0000000	.111111111	59	34 81	7.6811457	.016949153
10	1 00	3.1622777	.100000000	60	36 00	7.7459667	.016666667
11	1 21	3.3166248	.090909091	61	37 21	7.8102497	.016393443
12	1 44	3.4641016	.083333333	62	38 44	7.8740079	.016129032
13	1 69	3.6055513	.076923077	63	39 69	7.9372539	.015873016
14	1 96	3.7416574	.071428571	64	40 96	8.0000000	.015625000
15	2 25	3.8729833	.066666667	65	42 25	8.0622577	.015384615
16	2 56	4.0000000	.062500000	66	43 56	8.1240384	.015151515
17	2 89	4.1231056	.058823529	67	44 89	8.1853528	.014925373
18	3 24	4.2426407	.055555556	68	46 24	8.2462113	.014705882
19	3 61	4.3588989	.052631579	69	47 61	8.3066239	.014492754
20	4 00	4.4721360	.050000000	70	49 00	8.3666003	.014285714
21	4 41	4.5825757	.047619048	71	50 41	8.4261498	.014084507
22	4 84	4.6904158	.045454545	72	51 84	8.4852814	.013888889
23	5 29	4.7958315	.043478261	73	53 29	8.5440037	.013698630
24	5 76	4.8989795	.041666667	74	54 76	8.6023253	.013513514
25	6 25	5.0000000	.040000000	75	56 25	8.6602540	.013333333
26	6 76	5.0990195	.038461538	76	57 76	8.7177979	.013157895
27	7 29	5.1961524	.037037037	77	59 29	8.7749644	.012987013
28	7 84	5.2915026	.035714286	78	60 84	8.8317609	.012820513
29	8 41	5.3851648	.034482759	79	62 41	8.8881944	.012658228
30	9 00	5.4772256	.033333333	80	64 00	8.9442719	.012500000
31	9 61	5.5677644	.032258065	81	65 61	9.0000000	.012345679
32	10 24	5.6568542	.031250000	82	67 24	9.0553851	.012195122
33	10 89	5.7445626	.030303030	83	68 89	9.1104336	.012048193
34	11 56	5.8309519	.029411765	84	70 56	9.1651514	.011904762
35	12 25	5.9160798	.028571429	85	72 25	9.2195445	.011764706
36	12 96	6.0000000	.027777778	86	73 96	9.2736185	.011627907
37	13 69	6.0827625	.027027027	87	75 69	9.3273791	.011494253
38	14 44	6.1644140	.026315789	88	77 44	9.3808315	.011363636
39	15 21	6.2449980	.025641026	89	79 21	9.4339811	.011235955
40	16 00	6.3245553	.025000000	90	81 00	9.4868330	.011111111
41	16 81	6.4031242	.024390244	91	82 81	9.5393920	.010989011
42	17 64	6.4807407	.023809524	92	84 64	9.5916630	.010869565
43	18 49	6.5574385	.023255814	93	86 49	9.6436508	.010752688
44	19 36	6.6332496	.022727273	94	88 36	9.6953597	.010638298
45	20 25	6.7082039	.022222222	95	90 25	9.7467943	.010526316
46	21 16	6.7823300	.021739130	96	92 16	9.7979590	.010416667
47	22 09	6.8556546	.021276596	97	94 09	9.8488578	.010309278
48	23 04	6.9282032	.020833333	98	96 04	9.8994949	.010204082
49	24 01	7.0000000	.020408163	99	98 01	9.9498744	.010101010
50	25 00	7.0710678	.020000000	100	1 00 00	10.0000000	.010000000

[1] The following ten tables from Chaddock, R. E., and Croxton, F. E., *Exercises in Statistical Method*, by courtesy of Houghton Mifflin Company.

No.	Square	Square Root	Reciprocal .00
101	1 02 01	10.0498756	9900990
102	1 04 04	10.0995049	9803922
103	1 06 09	10.1488916	9708738
104	1 08 16	10.1980390	9615385
105	1 10 25	10.2469508	9523810
106	1 12 36	10.2956301	9433962
107	1 14 49	10.3440804	9345794
108	1 16 64	10.3923048	9259259
109	1 18 81	10.4403065	9174312
110	1 21 00	10.4880885	9090909
111	1 23 21	10.5356538	9009009
112	1 25 44	10.5830052	8928571
113	1 27 69	10.6301458	8849558
114	1 29 96	10.6770783	8771930
115	1 32 25	10.7238053	8695652
116	1 34 56	10.7703296	8620690
117	1 36 89	10.8166538	8547009
118	1 39 24	10.8627805	8474576
119	1 41 61	10.9087121	8403361
120	1 44 00	10.9544512	8333333
121	1 46 41	11.0000000	8264463
122	1 48 84	11.0453610	8196721
123	1 51 29	11.0905365	8130081
124	1 53 76	11.1355287	8064516
125	1 56 25	11.1803399	8000000
126	1 58 76	11.2249722	7936508
127	1 61 29	11.2694277	7874016
128	1 63 84	11.3137085	7812500
129	1 66 41	11.3578167	7751938
130	1 69 00	11.4017543	7692308
131	1 71 61	11.4455231	7633588
132	1 74 24	11.4891253	7575758
133	1 76 89	11.5325626	7518797
134	1 79 56	11.5758369	7462687
135	1 82 25	11.6189500	7407407
136	1 84 96	11.6619038	7352941
137	1 87 69	11.7046999	7299270
138	1 90 44	11.7473401	7246377
139	1 93 21	11.7898261	7194245
140	1 96 00	11.8321596	7142857
141	1 98 81	11.8743422	7092199
142	2 01 64	11.9163753	7042254
143	2 04 49	11.9582607	6993007
144	2 07 36	12.0000000	6944444
145	2 10 25	12.0415946	6896552
146	2 13 16	12.0830460	6849315
147	2 16 09	12.1243557	6802721
148	2 19 04	12.1655251	6756757
149	2 22 01	12.2065556	6711409
150	2 25 00	12.2474487	6666667
151	2 28 01	12.2882057	6622517
152	2 31 04	12.3288280	6578947
153	2 34 09	12.3693169	6535948
154	2 37 16	12.4096736	6493506
155	2 40 25	12.4498996	6451613
156	2 43 36	12.4899960	6410256
157	2 46 49	12.5299641	6369427
158	2 49 64	12.5698051	6329114
159	2 52 81	12.6095202	6289308
160	2 56 00	12.6491106	6250000
161	2 59 21	12.6885775	6211180
162	2 62 44	12.7279221	6172840
163	2 65 69	12.7671453	6134969
164	2 68 96	12.8062485	6097561
165	2 72 25	12.8452326	6060606
166	2 75 56	12.8840987	6024096
167	2 78 89	12.9228480	5988024
168	2 82 24	12.9614814	5952381
169	2 85 61	13.0000000	5917160
170	2 89 00	13.0384048	5882353
171	2 92 41	13.0766968	5847953
172	2 95 84	13.1148770	5813953
173	2 99 29	13.1529464	5780347
174	3 02 76	13.1909060	5747126
175	3 06 25	13.2287566	5714286
176	3 09 76	13.2664992	5681818
177	3 13 29	13.3041347	5649718
178	3 16 84	13.3416641	5617978
179	3 20 41	13.3790882	5586592
180	3 24 00	13.4164079	5555556
181	3 27 61	13.4536240	5524862
182	3 31 24	13.4907376	5494505
183	3 34 89	13.5277493	5464481
184	3 38 56	13.5646600	5434783
185	3 42 25	13.6014705	5405405
186	3 45 96	13.6381817	5376344
187	3 49 69	13.6747943	5347594
188	3 53 44	13.7113092	5319149
189	3 57 21	13.7477271	5291005
190	3 61 00	13.7840488	5263158
191	3 64 81	13.8202750	5235602
192	3 68 64	13.8564065	5208333
193	3 72 49	13.8924440	5181347
194	3 76 36	13.9283883	5154639
195	3 80 25	13.9642400	5128205
196	3 84 16	14.0000000	5102041
197	3 88 09	14.0356688	5076142
198	3 92 04	14.0712473	5050505
199	3 96 01	14.1067360	5025126
200	4 00 00	14.1421356	5000000

No.	Square	Square Root	Reciprocal .00
201	4 04 01	14.1774469	4975124
202	4 08 04	14.2126704	4950495
203	4 12 09	14.2478068	4926108
204	4 16 16	14.2828569	4901961
205	4 20 25	14.3178211	4878049
206	4 24 36	14.3527001	4854369
207	4 28 49	14.3874946	4830918
208	4 32 64	14.4222051	4807692
209	4 36 81	14.4568323	4784689
210	4 41 00	14.4913767	4761905
211	4 45 21	14.5258390	4739336
212	4 49 44	14.5602198	4716981
213	4 53 69	14.5945195	4694836
214	4 57 96	14.6287388	4672897
215	4 62 25	14.6628783	4651163
216	4 66 56	14.6969385	4629630
217	4 70 89	14.7309199	4608295
218	4 75 24	14.7648231	4587156
219	4 79 61	14.7986486	4566210
220	4 84 00	14.8323970	4545455
221	4 88 41	14.8660687	4524887
222	4 92 84	14.8996644	4504505
223	4 97 29	14.9331845	4484305
224	5 01 76	14.9666295	4464286
225	5 06 25	15.0000000	4444444
226	5 10 76	15.0332964	4424779
227	5 15 29	15.0665192	4405286
228	5 19 84	15.0996689	4385965
229	5 24 41	15.1327460	4366812
230	5 29 00	15.1657509	4347826
231	5 33 61	15.1986842	4329004
232	5 38 24	15.2315462	4310345
233	5 42 89	15.2643375	4291845
234	5 47 56	15.2970585	4273504
235	5 52 25	15.3297097	4255319
236	5 56 96	15.3622915	4237288
237	5 61 69	15.3948043	4219409
238	5 66 44	15.4272486	4201681
239	5 71 21	15.4596248	4184100
240	5 76 00	15.4919334	4166667
241	5 80 81	15.5241747	4149378
242	5 85 64	15.5563492	4132231
243	5 90 49	15.5884573	4115226
244	5 95 36	15.6204994	4098361
245	6 00 25	15.6524758	4081633
246	6 05 16	15.6843871	4065041
247	6 10 09	15.7162336	4048583
248	6 15 04	15.7480157	4032258
249	6 20 01	15.7797338	4016064
250	6 25 00	15.8113883	4000000
251	6 30 01	15.8429795	3984064
252	6 35 04	15.8745079	3968254
253	6 40 09	15.9059737	3952569
254	6 45 16	15.9373775	3937008
255	6 50 25	15.9687194	3921569
256	6 55 36	16.0000000	3906250
257	6 60 49	16.0312195	3891051
258	6 65 64	16.0623784	3875969
259	6 70 81	16.0934769	3861004
260	6 76 00	16.1245155	3846154
261	6 81 21	16.1554944	3831418
262	6 86 44	16.1864141	3816794
263	6 91 69	16.2172747	3802281
264	6 96 96	16.2480768	3787879
265	7 02 25	16.2788206	3773585
266	7 07 56	16.3095064	3759398
267	7 12 89	16.3401346	3745318
268	7 18 24	16.3707055	3731343
269	7 23 61	16.4012195	3717472
270	7 29 00	16.4316767	3703704
271	7 34 41	16.4620776	3690037
272	7 39 84	16.4924225	3676471
273	7 45 29	16.5227116	3663004
274	7 50 76	16.5529454	3649635
275	7 56 25	16.5831240	3636364
276	7 61 76	16.6132477	3623188
277	7 67 29	16.6433170	3610108
278	7 72 84	16.6733320	3597122
279	7 78 41	16.7032931	3584229
280	7 84 00	16.7332005	3571429
281	7 89 61	16.7630546	3558719
282	7 95 24	16.7928556	3546099
283	8 00 89	16.8226038	3533569
284	8 06 56	16.8522995	3521127
285	8 12 25	16.8819430	3508772
286	8 17 96	16.9115345	3496503
287	8 23 69	16.9410743	3484321
288	8 29 44	16.9705627	3472222
289	8 35 21	17.0000000	3460208
290	8 41 00	17.0293864	3448276
291	8 46 81	17.0587221	3436426
292	8 52 64	17.0880075	3424658
293	8 58 49	17.1172428	3412969
294	8 64 36	17.1464282	3401361
295	8 70 25	17.1755640	3389831
296	8 76 16	17.2046505	3378378
297	8 82 09	17.2336879	3367003
298	8 88 04	17.2626765	3355705
299	8 94 01	17.2916165	3344482
300	9 00 00	17.3205081	3333333

No.	Square	Square Root	Reciprocal .00
301	9 06 01	17.3493516	3322259
302	9 12 04	17.3781472	3311258
303	9 18 09	17.4068952	3300330
304	9 24 16	17.4355958	3289474
305	9 30 25	17.4642492	3278689
306	9 36 36	17.4928557	3267974
307	9 42 49	17.5214155	3257329
308	9 48 64	17.5499288	3246753
309	9 54 81	17.5783958	3236246
310	9 61 00	17.6068169	3225806
311	9 67 21	17.6351921	3215434
312	9 73 44	17.6635217	3205128
313	9 79 69	17.6918060	3194888
314	9 85 96	17.7200451	3184713
315	9 92 25	17.7482393	3174603
316	9 98 56	17.7763888	3164557
317	10 04 89	17.8044938	3154574
318	10 11 24	17.8325545	3144654
319	10 17 61	17.8605711	3134796
320	10 24 00	17.8885438	3125000
321	10 30 41	17.9164729	3115265
322	10 36 84	17.9443584	3105590
323	10 43 29	17.9722008	3095975
324	10 49 76	18.0000000	3086420
325	10 56 25	18.0277564	3076923
326	10 62 76	18.0554701	3067485
327	10 69 29	18.0831413	3058104
328	10 75 84	18.1107703	3048780
329	10 82 41	18.1383571	3039514
330	10 89 00	18.1659021	3030303
331	10 95 61	18.1934054	3021148
332	11 02 24	18.2208672	3012048
333	11 08 89	18.2482876	3003003
334	11 15 56	18.2756669	2994012
335	11 22 25	18.3030052	2985075
336	11 28 96	18.3303028	2976190
337	11 35 69	18.3575598	2967359
338	11 42 44	18.3847763	2958580
339	11 49 21	18.4119526	2949853
340	11 56 00	18.4390889	2941176
341	11 62 81	18.4661853	2932551
342	11 69 64	18.4932420	2923977
343	11 76 49	18.5202592	2915452
344	11 83 36	18.5472370	2906977
345	11 90 25	18.5741756	2898551
346	11 97 16	18.6010752	2890173
347	12 04 09	18.6279360	2881844
348	12 11 04	18.6547581	2873563
349	12 18 01	18.6815417	2865330
350	12 25 00	18.7082869	2857143

No.	Square	Square Root	Reciprocal .00
351	12 32 01	18.7349940	2849003
352	12 39 04	18.7616630	2840909
353	12 46 09	18.7882942	2832861
354	12 53 16	18.8148877	2824859
355	12 60 25	18.8414437	2816901
356	12 67 36	18.8679623	2808989
357	12 74 49	18.8944436	2801120
358	12 81 64	18.9208879	2793296
359	12 88 81	18.9472953	2785515
360	12 96 00	18.9736660	2777778
361	13 03 21	19.0000000	2770083
362	13 10 44	19.0262976	2762431
363	13 17 69	19.0525589	2754821
364	13 24 96	19.0787840	2747253
365	13 32 25	19.1049732	2739726
366	13 39 56	19.1311265	2732240
367	13 46 89	19.1572441	2724796
368	13 54 24	19.1833261	2717391
369	13 61 61	19.2093727	2710027
370	13 69 00	19.2353841	2702703
371	13 76 41	19.2613603	2695418
372	13 83 84	19.2873015	2688172
373	13 91 29	19.3132079	2680965
374	13 98 76	19.3390796	2673797
375	14 06 25	19.3649167	2666667
376	14 13 76	19.3907194	2659574
377	14 21 29	19.4164878	2652520
378	14 28 84	19.4422221	2645503
379	14 36 41	19.4679223	2638522
380	14 44 00	19.4935887	2631579
381	14 51 61	19.5192213	2624672
382	14 59 24	19.5448203	2617801
383	14 66 89	19.5703858	2610966
384	14 74 56	19.5959179	2604167
385	14 82 25	19.6214169	2597403
386	14 89 96	19.6468827	2590674
387	14 97 69	19.6723156	2583979
388	15 05 44	19.6977156	2577320
389	15 13 21	19.7230829	2570694
390	15 21 00	19.7484177	2564103
391	15 28 81	19.7737199	2557545
392	15 36 64	19.7989899	2551020
393	15 44 49	19.8242276	2544529
394	15 52 36	19.8494332	2538071
395	15 60 25	19.8746069	2531646
396	15 68 16	19.8997487	2525253
397	15 76 09	19.9248588	2518892
398	15 84 04	19.9499373	2512563
399	15 92 01	19.9749844	2506266
400	16 00 00	20.0000000	2500000

No.	Square	Square Root	Reciprocal .00
401	16 08 01	20.0249844	2493766
402	16 16 04	20.0499377	2487562
403	16 24 09	20.0748599	2481390
404	16 32 16	20.0997512	2475248
405	16 40 25	20.1246118	2469136
406	16 48 36	20.1494417	2463054
407	16 56 49	20.1742410	2457002
408	16 64 64	20.1990099	2450980
409	16 72 81	20.2237484	2444988
410	16 81 00	20.2484567	2439024
411	16 89 21	20.2731349	2433090
412	16 97 44	20.2977831	2427184
413	17 05 69	20.3224014	2421308
414	17 13 96	20.3469899	2415459
415	17 22 25	20.3715488	2409639
416	17 30 56	20.3960781	2403846
417	17 38 89	20.4205779	2398082
418	17 47 24	20.4450483	2392344
419	17 55 61	20.4694895	2386635
420	17 64 00	20.4939015	2380952
421	17 72 41	20.5182845	2375297
422	17 80 84	20.5426386	2369668
423	17 89 29	20.5669638	2364066
424	17 97 76	20.5912603	2358491
425	18 06 25	20.6155281	2352941
426	18 14 76	20.6397674	2347418
427	18 23 29	20.6639783	2341920
428	18 31 84	20.6881609	2336449
429	18 40 41	20.7123152	2331002
430	18 49 00	20.7364414	2325581
431	18 57 61	20.7605395	2320186
432	18 66 24	20.7846097	2314815
433	18 74 89	20.8086520	2309469
434	18 83 56	20.8326667	2304147
435	18 92 25	20.8566536	2298851
436	19 00 96	20.8806130	2293578
437	19 09 69	20.9045450	2288330
438	19 18 44	20.9284495	2283105
439	19 27 21	20.9523268	2277904
440	19 36 00	20.9761770	2272727
441	19 44 81	21.0000000	2267574
442	19 53 64	21.0237960	2262443
443	19 62 49	21.0475652	2257336
444	19 71 36	21.0713075	2252252
445	19 80 25	21.0950231	2247191
446	19 89 16	21.1187121	2242152
447	19 98 09	21.1423745	2237136
448	20 07 04	21.1660105	2232143
449	20 16 01	21.1896201	2227171
450	20 25 00	21.2132034	2222222

No	Square	Square Root	Reciprocal .00
451	20 34 01	21.2367606	2217295
452	20 43 04	21.2602916	2212389
453	20 52 09	21.2837967	2207506
454	20 61 16	21.3072758	2202643
455	20 70 25	21.3307290	2197802
456	20 79 36	21.3541565	2192982
457	20 88 49	21.3775583	2188184
458	20 97 64	21.4009346	2183406
459	21 06 81	21.4242853	2178649
460	21 16 00	21.4476106	2173913
461	21 25 21	21.4709106	2169197
462	21 34 44	21.4941853	2164502
463	21 43 69	21.5174348	2159827
464	21 52 96	21.5406592	2155172
465	21 62 25	21.5638587	2150538
466	21 71 56	21.5870331	2145923
467	21 80 89	21.6101828	2141328
468	21 90 24	21.6333077	2136752
469	21 99 61	21.6564078	2132196
470	22 09 00	21.6794834	2127660
471	22 18 41	21.7025344	2123142
472	22 27 84	21.7255610	2118644
473	22 37 29	21.7485632	2114165
474	22 46 76	21.7715411	2109705
475	22 56 25	21.7944947	2105263
476	22 65 76	21.8174242	2100840
477	22 75 29	21.8403297	2096436
478	22 84 84	21.8632111	2092050
479	22 94 41	21.8860686	2087683
480	23 04 00	21.9089023	2083333
481	23 13 61	21.9317122	2079002
482	23 23 24	21.9544984	2074689
483	23 32 89	21.9772610	2070393
484	23 42 56	22.0000000	2066116
485	23 52 25	22.0227155	2061856
486	23 61 96	22.0454077	2057613
487	23 71 69	22.0680765	2053388
488	23 81 44	22.0907220	2049180
489	23 91 21	22.1133444	2044990
490	24 01 00	22.1359436	2040816
491	24 10 81	22.1585198	2036660
492	24 20 64	22.1810730	2032520
493	24 30 49	22.2036033	2028398
494	24 40 36	22.2261108	2024291
495	24 50 25	22.2485955	2020202
496	24 60 16	22.2710575	2016129
497	24 70 09	22.2934968	2012072
498	24 80 04	22.3159136	2008032
499	24 90 01	22.3383079	2004008
500	25 00 00	22.3606798	2000000

No.	Square	Square Root	Reciprocal .00
501	25 10 01	22.3830293	1996008
502	25 20 04	22.4053565	1992032
503	25 30 09	22.4276615	1988072
504	25 40 16	22.4499443	1984127
505	25 50 25	22.4722051	1980198
506	25 60 36	22.4944438	1976285
507	25 70 49	22.5166605	1972387
508	25 80 64	22.5388553	1968504
509	25 90 81	22.5610283	1964637
510	26 01 00	22.5831796	1960784
511	26 11 21	22.6053091	1956947
512	26 21 44	22.6274170	1953125
513	26 31 69	22.6495033	1949318
514	26 41 96	22.6715681	1945525
515	26 52 25	22.6936114	1941748
516	26 62 56	22.7156334	1937984
517	26 72 89	22.7376340	1934236
518	26 83 24	22.7596134	1930502
519	26 93 61	22.7815715	1926782
520	27 04 00	22.8035085	1923077
521	27 14 41	22.8254244	1919386
522	27 24 84	22.8473193	1915709
523	27 35 29	22.8691933	1912046
524	27 45 76	22.8910463	1908397
525	27 56 25	22.9128785	1904762
526	27 66 76	22.9346899	1901141
527	27 77 29	22.9564806	1897533
528	27 87 84	22.9782506	1893939
529	27 98 41	23.0000000	1890359
530	28 09 00	23.0217289	1886792
531	28 19 61	23.0434372	1883239
532	28 30 24	23.0651252	1879699
533	28 40 89	23.0867928	1876173
534	28 51 56	23.1084400	1872659
535	28 62 25	23.1300670	1869159
536	28 72 96	23.1516738	1865672
537	28 83 69	23.1732605	1862197
538	28 94 44	23.1948270	1858736
539	29 05 21	23.2163735	1855288
540	29 16 00	23.2379001	1851852
541	29 26 81	23.2594067	1848429
542	29 37 64	23.2808935	1845018
543	29 48 49	23.3023604	1841621
544	29 59 36	23.3238076	1838235
545	29 70 25	23.3452351	1834862
546	29 81 16	23.3666429	1831502
547	29 92 09	23.3880311	1828154
548	30 03 04	23.4093998	1824818
549	30 14 01	23.4307490	1821494
550	30 25 00	23.4520788	1818182

No.	Square	Square Root	Reciprocal .00
551	30 36 01	23.4733892	1814882
552	30 47 04	23.4946802	1811594
553	30 58 09	23.5159520	1808318
554	30 69 16	23.5372046	1805054
555	30 80 25	23.5584380	1801802
556	30 91 36	23.5796522	1798561
557	31 02 49	23.6008474	1795332
558	31 13 64	23.6220236	1792115
559	31 24 81	23.6431808	1788909
560	31 36 00	23.6643191	1785714
561	31 47 21	23.6854386	1782531
562	31 58 44	23.7065392	1779359
563	31 69 69	23.7276210	1776199
564	31 80 96	23.7486842	1773050
565	31 92 25	23.7697286	1769912
566	32 03 56	23.7907545	1766784
567	32 14 89	23.8117618	1763668
568	32 26 24	23.8327506	1760563
569	32 37 61	23.8537209	1757469
570	32 49 00	23.8746728	1754386
571	32 60 41	23.8956063	1751313
572	32 71 84	23.9165215	1748252
573	32 83 29	23.9374184	1745201
574	32 94 76	23.9582971	1742160
575	33 06 25	23.9791576	1739130
576	33 17 76	24.0000000	1736111
577	33 29 29	24.0208243	1733102
578	33 40 84	24.0416306	1730104
579	33 52 41	24.0624188	1727116
580	33 64 00	24.0831891	1724138
581	33 75 61	24.1039416	1721170
582	33 87 24	24.1246762	1718213
583	33 98 89	24.1453929	1715266
584	34 10 56	24.1660919	1712329
585	34 22 25	24.1867732	1709402
586	34 33 96	24.2074369	1706485
587	34 45 69	24.2280829	1703578
588	34 57 44	24.2487113	1700680
589	34 69 21	24.2693222	1697793
590	34 81 00	24.2899156	1694915
591	34 92 81	24.3104916	1692047
592	35 04 64	24.3310501	1689189
593	35 16 49	24.3515913	1686341
594	35 28 36	24.3721152	1683502
595	35 40 25	24.3926218	1680672
596	35 52 16	24.4131112	1677852
597	35 64 09	24.4335834	1675042
598	35 76 04	24.4540385	1672241
599	35 88 01	24.4744765	1669449
600	36 00 00	24.4948974	1666667

No.	Square	Square Root	Reciprocal .00
601	36 12 01	24.5153013	1663894
602	36 24 04	24.5356883	1661130
603	36 36 09	24.5560583	1658375
604	36 48 16	24.5764115	1655629
605	36 60 25	24.5967478	1652893
606	36 72 36	24.6170673	1650165
607	36 84 49	24.6373700	1647446
608	36 96 64	24.6576560	1644737
609	37 08 81	24.6779254	1642036
610	37 21 00	24.6981781	1639344
611	37 33 21	24.7184142	1636661
612	37 45 44	24.7386338	1633987
613	37 57 69	24.7588368	1631321
614	37 69 96	24.7790234	1628664
615	37 82 25	24.7991935	1626016
616	37 94 56	24.8193473	1623377
617	38 06 89	24.8394847	1620746
618	38 19 24	24.8596058	1618123
619	38 31 61	24.8797106	1615509
620	38 44 00	24.8997992	1612903
621	38 56 41	24.9198716	1610306
622	38 68 84	24.9399278	1607717
623	38 81 29	24.9599679	1605136
624	38 93 76	24.9799920	1602564
625	39 06 25	25.0000000	1600000
626	39 18 76	25.0199920	1597444
627	39 31 29	25.0399681	1594896
628	39 43 84	25.0599282	1592357
629	39 56 41	25.0798724	1589825
630	39 69 00	25.0998008	1587302
631	39 81 61	25.1197134	1584786
632	39 94 24	25.1396102	1582278
633	40 06 89	25.1594913	1579779
634	40 19 56	25.1793566	1577287
635	40 32 25	25.1992063	1574803
636	40 44 96	25.2190404	1572327
637	40 57 69	25.2388589	1569859
638	40 70 44	25.2586619	1567398
639	40 83 21	25.2784493	1564945
640	40 96 00	25.2982213	1562500
641	41 08 81	25.3179778	1560062
642	41 21 64	25.3377189	1557632
643	41 34 49	25.3574447	1555210
644	41 47 36	25.3771551	1552795
645	41 60 25	25.3968502	1550388
646	41 73 16	25.4165301	1547988
647	41 86 09	25.4361947	1545595
648	41 99 04	25.4558441	1543210
649	42 12 01	25.4754784	1540832
650	42 25 00	25.4950976	1538462

No.	Square	Square Root	Reciprocal .00
651	42 38 01	25.5147016	1536098
652	42 51 04	25.5342907	1533742
653	42 64 09	25.5538647	1531394
654	42 77 16	25.5734237	1529052
655	42 90 25	25.5929678	1526718
656	43 03 36	25.6124969	1524390
657	43 16 49	25.6320112	1522070
658	43 29 64	25.6515107	1519757
659	43 42 81	25.6709953	1517451
660	43 56 00	25.6904652	1515152
661	43 69 21	25.7099203	1512859
662	43 82 44	25.7293607	1510574
663	43 95 69	25.7487864	1508296
664	44 08 96	25.7681975	1506024
665	44 22 25	25.7875939	1503759
666	44 35 56	25.8069758	1501502
667	44 48 89	25.8263431	1499250
668	44 62 24	25.8456960	1497006
669	44 75 61	25.8650343	1494768
670	44 89 00	25.8843582	1492537
671	45 02 41	25.9036677	1490313
672	45 15 84	25.9229628	1488095
673	45 29 29	25.9422435	1485884
674	45 42 76	25.9615100	1483680
675	45 56 25	25.9807621	1481481
676	45 69 76	26.0000000	1479290
677	45 83 29	26.0192237	1477105
678	45 96 84	26.0384331	1474926
679	46 10 41	26.0576284	1472754
680	46 24 00	26.0768096	1470588
681	46 37 61	26.0959767	1468429
682	46 51 24	26.1151297	1466276
683	46 64 89	26.1342687	1464129
684	46 78 56	26.1533937	1461988
685	46 92 25	26.1725047	1459854
686	47 05 96	26.1916017	1457726
687	47 19 69	26.2106848	1455604
688	47 33 44	26.2297541	1453488
689	47 47 21	26.2488095	1451379
690	47 61 00	26.2678511	1449275
691	47 74 81	26.2868789	1447178
692	47 88 64	26.3058929	1445087
693	48 02 49	26.3248932	1443001
694	48 16 36	26.3438797	1440922
695	48 30 25	26.3628527	1438849
696	48 44 16	26.3818119	1436782
697	48 58 09	26.4007576	1434720
698	48 72 04	26.4196896	1432665
699	48 86 01	26.4386081	1430615
700	49 00 00	26.4575131	1428571

No.	Square	Square Root	Reciprocal .00	No.	Square	Square Root	Reciprocal .00
701	49 14 01	26.4764046	1426534	751	56 40 01	27.4043792	1331558
702	49 28 04	26.4952826	1424501	752	56 55 04	27.4226184	1329787
703	49 42 09	26.5141472	1422475	753	56 70 09	27.4408455	1328021
704	49 56 16	26.5329983	1420455	754	56 85 16	27.4590604	1326260
705	49 70 25	26.5518361	1418440	755	57 00 25	27.4772633	1324503
706	49 84 36	26.5706605	1416431	756	57 15 36	27.4954542	1322751
707	49 98 49	26.5894716	1414427	757	57 30 49	27.5136330	1321004
708	50 12 64	26.6082694	1412429	758	57 45 64	27.5317998	1319261
709	50 26 81	26.6270539	1410437	759	57 60 81	27.5499546	1317523
710	50 41 00	26.6458252	1408451	760	57 76 00	27.5680975	1315789
711	50 55 21	26.6645833	1406470	761	57 91 21	27.5862284	1314060
712	50 69 44	26.6833281	1404494	762	58 06 44	27.6043475	1312336
713	50 83 69	26.7020598	1402525	763	58 21 69	27.6224546	1310616
714	50 97 96	26.7207784	1400560	764	58 36 96	27.6405499	1308901
715	51 12 25	26.7394839	1398601	765	58 52 25	27.6586334	1307190
716	51 26 56	26.7581763	1396648	766	58 67 56	27.6767050	1305483
717	51 40 89	26.7768557	1394700	767	58 82 89	27.6947648	1303781
718	51 55 24	26.7955220	1392758	768	58 98 24	27.7128129	1302083
719	51 69 61	26.8141754	1390821	769	59 13 61	27.7308492	1300390
720	51 84 00	26.8328157	1388889	770	59 29 00	27.7488739	1298701
721	51 98 41	26.8514432	1386963	771	59 44 41	27.7668868	1297017
722	52 12 84	26.8700577	1385042	772	59 59 84	27.7848880	1295337
723	52 27 29	26.8886593	1383126	773	59 75 29	27.8028775	1293661
724	52 41 76	26.9072481	1381215	774	59 90 76	27.8208555	1291990
725	52 56 25	26.9258240	1379310	775	60 06 25	27.8388218	1290323
726	52 70 76	26.9443872	1377410	776	60 21 06	27.8567766	1288660
727	52 85 29	26.9629375	1375516	777	60 37 29	27.8747197	1287001
728	52 99 84	26.9814751	1373626	778	60 52 84	27.8926514	1285347
729	53 14 41	27.0000000	1371742	779	60 68 41	27.9105715	1283697
730	53 29 00	27.0185122	1369863	780	60 84 00	27.9284801	1282051
731	53 43 61	27.0370117	1367989	781	60 99 61	27.9463772	1280410
732	53 58 24	27.0554985	1366120	782	61 15 24	27.9642629	1278772
733	53 72 89	27.0739727	1364256	783	61 30 89	27.9821372	1277139
734	53 87 56	27.0924344	1362398	784	61 46 56	28.0000000	1275510
735	54 02 25	27.1108834	1360544	785	61 62 25	28.0178515	1273885
736	54 16 96	27.1293199	1358696	786	61 77 96	28.0356915	1272265
737	54 31 69	27.1477439	1356852	787	61 93 69	28.0535203	1270648
738	54 46 44	27.1661554	1355014	788	62 09 44	28.0713377	1269036
739	54 61 21	27.1845544	1353180	789	62 25 21	28.0891438	1267427
740	54 76 00	27.2029410	1351351	790	62 41 00	28.1069386	1265823
741	54 90 81	27.2213152	1349528	791	62 56 81	28.1247222	1264223
742	55 05 64	27.2396769	1347709	792	62 72 64	28.1424946	1262626
743	55 20 49	27.2580263	1345895	793	62 88 49	28.1602557	1261034
744	55 35 36	27.2763634	1344086	794	63 04 36	28.1780056	1259446
745	55 50 25	27.2946881	1342282	795	63 20 25	28.1957444	1257862
746	55 65 16	27.3130006	1340483	796	63 36 16	28.2134720	1256281
747	55 80 09	27.3313007	1338688	797	63 52 09	28.2311884	1254705
748	55 95 04	27.3495887	1336898	798	63 68 04	28.2488938	1253133
749	56 10 01	27.3678644	1335113	799	63 84 01	28.2665881	1251564
750	56 25 00	27.3861279	1333333	800	64 00 00	28.2842712	1250000

No.	Square	Square Root	Reciprocal .00
801	64 16 01	28.3019434	1248439
802	64 32 04	28.3196045	1246883
803	64 48 09	28.3372546	1245330
804	64 64 16	28.3548938	1243781
805	64 80 25	28.3725219	1242236
806	64 96 36	28.3901391	1240695
807	65 12 49	28.4077454	1239157
808	65 28 64	28.4253408	1237624
809	65 44 81	28.4429253	1236094
810	65 61 00	28.4604989	1234568
811	65 77 21	28.4780617	1233046
812	65 93 44	28.4956137	1231527
813	66 09 69	28.5131549	1230012
814	66 25 96	28.5306852	1228501
815	66 42 25	28.5482048	1226994
816	66 58 56	28.5657137	1225490
817	66 74 89	28.5832119	1223990
818	66 91 24	28.6006993	1222494
819	67 07 61	28.6181760	1221001
820	67 24 00	28.6356421	1219512
821	67 40 41	28.6530976	1218027
822	67 56 84	28.6705424	1216545
823	67 73 29	28.6879766	1215067
824	67 89 76	28.7054002	1213592
825	68 06 25	28.7228132	1212121
826	68 22 76	28.7402157	1210654
827	68 39 29	28.7576077	1209190
828	68 55 84	28.7749891	1207729
829	68 72 41	28.7923601	1206273
830	68 89 00	28.8097206	1204819
831	69 05 61	28.8270706	1203369
832	69 22 24	28.8444102	1201923
833	69 38 89	28.8617394	1200480
834	69 55 56	28.8790582	1199041
835	69 72 25	28.8963666	1197605
836	69 88 96	28.9136646	1196172
837	70 05 69	28.9309523	1194743
838	70 22 44	28.9482297	1193317
839	70 39 21	28.9654967	1191895
840	70 56 00	28.9827535	1190476
841	70 72 81	29.0000000	1189061
842	70 89 64	29.0172363	1187648
843	71 06 49	29.0344623	1186240
844	71 23 36	29.0516781	1184834
845	71 40 25	29.0688837	1183432
846	71 57 16	29.0860791	1182033
847	71 74 09	29.1032644	1180638
848	71 91 04	29.1204396	1179245
849	72 08 01	29.1376046	1177856
850	72 25 00	29.1547595	1176471

No.	Square	Square Root	Reciprocal 00
851	72 42 01	29.1719043	1175088
852	72 59 04	29.1890390	1173709
853	72 76 09	29.2061637	1172333
854	72 93 16	29.2232784	1170960
855	73 10 25	29.2403830	1169591
856	73 27 36	29.2574777	1168224
857	73 44 49	29.2745623	1166861
858	73 61 64	29.2916370	1165501
859	73 78 81	29.3087018	1164144
860	73 96 00	29.3257566	1162791
861	74 13 21	29.3428015	1161440
862	74 30 44	29.3598365	1160093
863	74 47 69	29.3768616	1158749
864	74 64 96	29.3938769	1157407
865	74 82 25	29.4108823	1156069
866	74 99 56	29.4278779	1154734
867	75 16 89	29.4448637	1153403
868	75 34 24	29.4618397	1152074
869	75 51 61	29.4788059	1150748
870	75 69 00	29.4957624	1149425
871	75 86 41	29.5127091	1148106
872	76 03 84	29.5296461	1146789
873	76 21 29	29.5465734	1145475
874	76 38 76	29.5634910	1144165
875	76 56 25	29.5803989	1142857
876	76 73 76	29.5972972	1141553
877	76 91 29	29.6141858	1140251
878	77 08 84	29.6310648	1138952
879	77 26 41	29.6479342	1137656
880	77 44 00	29.6647939	1136364
881	77 61 61	29.6816442	1135074
882	77 79 24	29.6984848	1133787
883	77 96 89	29.7153159	1132503
884	78 14 56	29.7321375	1131222
885	78 32 25	29.7489496	1129944
886	78 49 96	29.7657521	1128668
887	78 67 69	29.7825452	1127396
888	78 85 44	29.7993289	1126126
889	79 03 21	29.8161030	1124859
890	79 21 00	29.8328678	1123596
891	79 38 81	29.8496231	1122334
892	79 56 64	29.8663690	1121076
893	79 74 49	29.8831056	1119821
894	79 92 36	29.8998328	1118568
895	80 10 25	29.9165506	1117318
896	80 28 16	29.9332591	1116071
897	80 46 09	29.9499583	1114827
898	80 64 04	29.9666481	1113586
899	80 82 01	29.9833287	1112347
900	81 00 00	30.0000000	1111111

No.	Square	Square Root	Reciprocal .00
901	81 18 01	30.0166620	1109878
902	81 36 04	30.0333148	1108647
903	81 54 09	30.0499584	1107420
904	81 72 16	30.0665928	1106195
905	81 90 25	30.0832179	1104972
906	82 08 36	30.0998339	1103753
907	82 26 49	30.1164407	1102536
908	82 44 64	30.1330383	1101322
909	82 62 81	30.1496269	1100110
910	82 81 00	30.1662063	1098901
911	82 99 21	30.1827765	1097695
912	83 17 44	30.1993377	1096491
913	83 35 69	30.2158899	1095290
914	83 53 96	30.2324329	1094092
915	83 72 25	30.2489669	1092896
916	83 90 56	30.2654919	1091703
917	84 08 89	30.2820079	1090513
918	84 27 24	30.2985148	1089325
919	84 45 61	30.3150128	1088139
920	84 64 00	30.3315018	1086957
921	84 82 41	30.3479818	1085776
922	85 00 84	30.3644529	1084599
923	85 19 29	30.3809151	1083424
924	85 37 76	30.3973683	1082251
925	85 56 25	30.4138127	1081081
926	85 74 76	30.4302481	1079914
927	85 93 29	30.4466747	1078749
928	86 11 84	30.4630924	1077586
929	86 30 41	30.4795013	1076426
930	86 49 00	30.4959014	1075269
931	86 67 61	30.5122926	1074114
932	86 86 24	30.5286750	1072961
933	87 04 89	30.5450487	1071811
934	87 23 56	30.5614136	1070664
935	87 42 25	30.5777697	1069519
936	87 60 96	30.5941171	1068376
937	87 79 69	30.6104557	1067236
938	87 98 44	30.6267857	1066098
939	88 17 21	30.6431069	1064963
940	88 36 00	30.6594194	1063830
941	88 54 81	30.6757233	1062699
942	88 73 64	30.6920185	1061571
943	88 92 49	30.7083051	1060445
944	89 11 36	30.7245830	1059322
945	89 30 25	30.7408523	1058201
946	89 49 16	30.7571130	1057082
947	89 68 09	30.7733651	1055966
948	89 87 04	30.7896086	1054852
949	90 06 01	30.8058436	1053741
950	90 25 00	30.8220700	1052632
951	90 44 01	30.8382879	1051525
952	90 63 04	30.8544972	1050420
953	90 82 09	30.8706981	1049318
954	91 01 16	30.8868904	1048218
955	91 20 25	30.9030743	1047120
956	91 39 36	30.9192497	1046025
957	91 58 49	30.9354166	1044932
958	91 77 64	30.9515751	1043841
959	91 96 81	30.9677251	1042753
960	92 16 00	30.9838668	1041667
961	92 35 21	31.0000000	1040583
962	92 54 44	31.0161248	1039501
963	92 73 69	31.0322413	1038422
964	92 92 96	31.0483494	1037344
965	93 12 25	31.0644491	1036269
966	93 31 56	31.0805405	1035197
967	93 50 89	31.0966236	1034126
968	93 70 24	31.1126984	1033058
969	93 89 61	31.1287648	1031992
970	94 09 00	31.1448230	1030928
971	94 28 41	31.1608729	1029866
972	94 47 84	31.1769145	1028807
973	94 67 29	31.1929479	1027749
974	94 86 76	31.2089731	1026694
975	95 06 25	31.2249900	1025641
976	95 25 76	31.2409987	1024590
977	95 45 29	31.2569992	1023541
978	95 64 84	31.2729915	1022495
979	95 84 41	31.2889757	1021450
980	96 04 00	31.3049517	1020408
981	96 23 61	31.3209195	1019368
982	96 43 24	31.3368792	1018330
983	96 62 89	31.3528308	1017294
984	96 82 56	31.3687743	1016260
985	97 02 25	31.3847097	1015228
986	97 21 96	31.4006369	1014199
987	97 41 69	31.4165561	1013171
988	97 61 44	31.4324673	1012146
989	97 81 21	31.4483704	1011122
990	98 01 00	31.4642654	1010101
991	98 20 81	31.4801525	1009082
992	98 40 64	31.4960315	1008065
993	98 60 49	31.5119025	1007049
994	98 80 36	31.5277655	1006036
995	99 00 25	31.5436206	1005025
996	99 20 16	31.5594677	1004016
997	99 40 09	31.5753068	1003009
998	99 60 04	31.5911380	1002004
999	99 80 01	31.6069613	1001001
1000	1 00 00 00	31.6227766	1000000

APPENDIX C

TABLE CXXV

Common Logarithms and Proportional Parts [1]

Numbers 100–150 Logs 00000–17869

N	0	1	2	3	4	5	6	7	8	9	P.P.			
100	00 000	00 043	00 087	00 130	00 173	00 217	00 260	00 303	00 346	00 389		43	42	41
101	432	475	518	561	604	647	689	732	775	817	1	4.3	4.2	4.1
102	860	903	945	988	01 030	01 072	01 115	01 157	01 199	01 242	2	8.6	8.4	8.2
103	01 284	01 326	01 368	01 410	452	494	536	578	620	662	3	12.9	12.6	12.3
104	703	745	787	828	870	912	953	995	02 036	02 078	4	17.2	16.8	16.4
105	02 119	02 160	02 202	02 243	02 284	02 325	02 366	02 407	449	490	5	21.5	21.0	20.5
106	531	572	612	653	694	735	776	816	857	898	6	25.8	25.2	24.6
107	938	979	03 019	03 060	03 100	03 141	03 181	03 222	03 262	03 302	7	30.1	29.4	28.7
108	03 342	03 383	423	463	503	543	583	623	663	703	8	34.4	33.6	32.8
109	743	782	822	862	902	941	981	04 021	04 060	04 100	9	38.7	37.8	36.9
110	04 139	04 179	04 218	04 258	04 297	04 336	04 376	04 415	04 454	04 493		40	39	38
111	532	571	610	650	689	727	766	805	844	883	1	4.0	3.9	3.8
112	922	961	999	05 038	05 077	05 115	05 154	05 192	05 231	05 269	2	8.0	7.8	7.6
113	05 308	05 346	05 385	423	461	500	538	576	614	652	3	12.0	11.7	11.4
114	690	729	767	805	843	881	918	956	994	06 032	4	16.0	15.6	15.2
115	06 070	06 108	06 145	06 183	06 221	06 258	06 296	06 333	06 371	408	5	20.0	19.5	19.0
116	446	483	521	558	595	633	670	707	744	781	6	24.0	23.4	22.8
117	819	856	893	930	967	07 004	07 041	07 078	07 115	07 151	7	28.0	27.3	26.6
118	07 188	07 225	07 262	07 298	07 335	372	408	445	482	518	8	32.0	31.2	30.4
119	555	591	628	664	700	737	773	809	846	882	9	36.0	35.1	34.2
120	07 918	07 954	07 990	08 027	08 063	08 099	08 135	08 171	08 207	08 243		37	36	35
121	08 279	08 314	08 350	386	422	458	493	529	565	600	1	3.7	3.6	3.5
122	636	672	707	743	778	814	849	884	920	955	2	7.4	7.2	7.0
123	991	09 026	09 061	09 096	09 132	09 167	09 202	09 237	09 272	09 307	3	11.1	10.8	10.5
124	09 342	377	412	447	482	517	552	587	621	656	4	14.8	14.4	14.0
125	691	726	760	795	830	864	899	934	968	10 003	5	18.5	18.0	17.5
126	10 037	10 072	10 106	10 140	10 175	10 209	10 243	10 278	10 312	346	6	22.2	21.6	21.0
127	380	415	449	483	517	551	585	619	653	687	7	25.9	25.2	24.5
128	721	755	789	823	857	890	924	958	992	11 025	8	29.6	28.8	28.0
129	11 059	11 093	11 126	11 160	11 193	11 227	11 261	11 294	11 327	361	9	33.3	32.4	31.5
130	11 394	11 428	11 461	11 494	11 528	11 561	11 594	11 628	11 661	11 694		34	33	32
131	727	760	793	826	860	893	926	959	992	12 024	1	3.4	3.3	3.2
132	12 057	12 090	12 123	12 156	12 189	12 222	12 254	12 287	12 320	352	2	6.8	6.6	6.4
133	385	418	450	483	516	548	581	613	646	678	3	10.2	9.9	9.6
134	710	743	775	808	840	872	905	937	969	13 001	4	13.6	13.2	12.8
135	13 033	13 066	13 098	13 130	13 162	13 194	13 226	13 258	13 290	322	5	17.0	16.5	16.0
136	354	386	418	450	481	513	545	577	609	640	6	20.4	19.8	19.2
137	672	704	735	767	799	830	862	893	925	956	7	23.8	23.1	22.4
138	988	14 019	14 051	14 082	14 114	14 145	14 176	14 208	14 239	14 270	8	27.2	26.4	25.6
139	14 301	333	364	395	426	457	489	520	551	582	9	30.6	29.7	28.8
140	14 613	14 644	14 675	14 706	14 737	14 768	14 799	14 829	14 860	14 891		31	30	29
141	922	953	983	15 014	15 045	15 076	15 106	15 137	15 168	15 198	1	3.1	3.0	2.9
142	15 229	15 259	15 290	320	351	381	412	442	473	503	2	6.2	6.0	5.8
143	534	564	594	625	655	685	715	746	776	806	3	9.3	9.0	8.7
144	836	866	897	927	957	987	16 017	16 047	16 077	16 107	4	12.4	12.0	11.6
145	16 137	16 167	16 197	16 227	16 256	16 286	316	346	376	406	5	15.5	15.0	14.5
146	435	465	495	524	554	584	613	643	673	702	6	18.6	18.0	17.4
147	732	761	791	820	850	879	909	938	967	997	7	21.7	21.0	20.3
148	17 026	17 056	17 085	17 114	17 143	17 173	17 202	17 231	17 260	17 289	8	24.8	24.0	23.2
149	319	348	377	406	435	464	493	522	551	580	9	27.9	27.0	26.1
150	17 609	17 638	17 667	17 696	17 725	17 754	17 782	17 811	17 840	17 869				
N	**0**	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**				

[1] Reprinted from *The Mathematics of Finance*, Houghton Mifflin Co., Boston, by permission of the publisher.

Numbers 150–200 Logs 17609–30298

N	0	1	2	3	4	5	6	7	8	9
150	17 609	17 638	17 667	17 696	17 725	17 754	17 782	17 811	17 840	17 869
151	898	926	955	984	18 013	18 041	18 070	18 099	18 127	18 156
152	18 184	18 213	18 241	18 270	298	327	355	384	412	441
153	469	498	526	554	583	611	639	667	696	724
154	752	780	808	837	865	893	921	949	977	19 005
155	19 033	19 061	19 089	19 117	19 145	19 173	19 201	19 229	19 257	285
156	312	340	368	396	424	451	479	507	535	562
157	590	618	645	673	700	728	756	783	811	838
158	866	893	921	948	976	20 003	20 030	20 058	20 085	20 112
159	20 140	20 167	20 194	20 222	20 249	276	303	330	358	385
160	20 412	20 439	20 466	20 493	20 520	20 548	20 575	20 602	20 629	20 656
161	683	710	737	763	790	817	844	871	898	925
162	952	978	21 005	21 032	21 059	21 085	21 112	21 139	21 165	21 192
163	21 219	21 245	272	299	325	352	378	405	431	458
164	484	511	537	564	590	617	643	669	696	722
165	748	775	801	827	854	880	906	932	958	985
166	22 011	22 037	22 063	22 089	22 115	22 141	22 167	22 194	22 220	22 246
167	272	298	324	350	376	401	427	453	479	505
168	531	557	583	608	634	660	686	712	737	763
169	789	814	840	866	891	917	943	968	994	23 019
170	23 045	23 070	23 096	23 121	23 147	23 172	23 198	23 223	23 249	23 274
171	300	325	350	376	401	426	452	477	502	528
172	553	578	603	629	654	679	704	729	754	779
173	805	830	855	880	905	930	955	980	24 005	24 030
174	24 055	24 080	24 105	24 130	24 155	24 180	24 204	24 229	254	279
175	304	329	353	378	403	428	452	477	502	527
176	551	576	601	625	650	674	699	724	748	773
177	797	822	846	871	895	920	944	969	993	25 018
178	25 042	25 066	25 091	25 115	25 139	25 164	25 188	25 212	25 237	261
179	285	310	334	358	382	406	431	455	479	503
180	25 527	25 551	25 575	25 600	25 624	25 648	25 672	25 696	25 720	25 744
181	768	792	816	840	864	888	912	935	959	983
182	26 007	26 031	26 055	26 079	26 102	26 126	26 150	26 174	26 198	26 221
183	245	269	293	316	340	364	387	411	435	458
184	482	505	529	553	576	600	623	647	670	694
185	717	741	764	788	811	834	858	881	905	928
186	951	975	998	27 021	27 045	27 068	27 091	27 114	27 138	27 161
187	27 184	27 207	27 231	254	277	300	323	346	370	393
188	416	439	462	485	508	531	554	577	600	623
189	646	669	692	715	738	761	784	807	830	852
190	27 875	27 898	27 921	27 944	27 967	27 989	28 012	28 035	28 058	28 081
191	28 103	28 126	28 149	28 171	28 194	28 217	240	262	285	307
192	330	353	375	398	421	443	466	488	511	533
193	556	578	601	623	646	668	691	713	735	758
194	780	803	825	847	870	892	914	937	959	981
195	29 003	29 026	29 048	29 070	29 092	29 115	29 137	29 159	29 181	29 203
196	226	248	270	292	314	336	358	380	403	425
197	447	469	491	513	535	557	579	601	623	645
198	667	688	710	732	754	776	798	820	842	863
199	885	907	929	951	973	994	30 016	30 038	30 060	30 081
200	30 103	30 125	30 146	30 168	30 190	30 211	30 233	30 255	30 276	30 298
N	0	1	2	3	4	5	6	7	8	9

P.P.

	29	28
1	2.9	2.8
2	5.8	5.6
3	8.7	8.4
4	11.6	11.2
5	14.5	14.0
6	17.4	16.8
7	20.3	19.6
8	23.2	22.4
9	26.1	25.2
	27	**26**
1	2.7	2.6
2	5.4	5.2
3	8.1	7.8
4	10.8	10.4
5	13.5	13.0
6	16.2	15 6
7	18.9	18.2
8	21.6	20.8
9	24.3	23.4
	25	**24**
1	2.5	2.4
2	5.0	4.8
3	7.5	7.2
4	10.0	9.6
5	12.5	12.0
6	15.0	14.4
7	17.5	16.8
8	20.0	19.2
9	22.5	21.6
	23	**22**
1	2.3	2.2
2	4.6	4.4
3	6.9	6.6
4	9.2	8.8
5	11.5	11.0
6	13.8	13.2
7	16.1	15.4
8	18.4	17.6
9	20.7	19.8
	21	
1	2.1	
2	4.2	
3	6.3	
4	8.4	
5	10.5	
6	12.6	
7	14.7	
8	16.8	
9	18.9	

Numbers 200–250 Logs 30103–39950

N	0	1	2	3	4	5	6	7	8	9
200	30 103	30 125	30 146	30 168	30 190	30 211	30 233	30 255	30 276	30 298
201	320	341	363	384	406	428	449	471	492	514
202	535	557	578	600	621	643	664	685	707	728
203	750	771	792	814	835	856	878	899	920	942
204	963	984	31 006	31 027	31 048	31 069	31 091	31 112	31 133	31 154
205	31 175	31 197	218	239	260	281	302	323	345	366
206	387	408	429	450	471	492	513	534	555	576
207	597	618	639	660	681	702	723	744	765	785
208	806	827	848	869	890	911	931	952	973	994
209	32 015	32 035	32 056	32 077	32 098	32 118	32 139	32 160	32 181	32 201
210	32 222	32 243	32 263	32 284	32 305	32 325	32 346	32 366	32 387	32 408
211	428	449	469	490	510	531	552	572	593	613
212	634	654	675	695	715	736	756	777	797	818
213	838	858	879	899	919	940	960	980	33 001	33 021
214	33 041	33 062	33 082	33 102	33 122	33 143	33 163	33 183	203	224
215	244	264	284	304	325	345	365	385	405	425
216	445	465	486	506	526	546	566	586	606	626
217	646	666	686	706	726	746	766	786	806	826
218	846	866	885	905	925	945	965	985	34 005	34 025
219	34 044	34 064	34 084	34 104	34 124	34 143	34 163	34 183	203	223
220	34 242	34 262	34 282	34 301	34 321	34 341	34 361	34 380	34 400	34 420
221	439	459	479	498	518	537	557	577	596	616
222	635	655	674	694	713	733	753	772	792	811
223	830	850	869	889	908	928	947	967	986	35 005
224	35 025	35 044	35 064	35 083	35 102	35 122	35 141	35 160	35 180	199
225	218	238	257	276	295	315	334	353	372	392
226	411	430	449	468	488	507	526	545	564	583
227	603	622	641	660	679	698	717	736	755	774
228	793	813	832	851	870	889	908	927	946	965
229	984	36 003	36 021	36 040	36 059	36 078	36 097	36 116	36 135	36 154
230	36 173	36 192	36 211	36 229	36 248	36 267	36 286	36 305	36 324	36 342
231	361	380	399	418	436	455	474	493	511	530
232	549	568	586	605	624	642	661	680	698	717
233	736	754	773	791	810	829	847	866	884	903
234	922	940	959	977	996	37 014	37 033	37 051	37 070	37 088
235	37 107	37 125	37 144	37 162	37 181	199	218	236	254	273
236	291	310	328	346	365	383	401	420	438	457
237	475	493	511	530	548	566	585	603	621	639
238	658	676	694	712	731	749	767	785	803	822
239	840	858	876	894	912	931	949	967	985	38 003
240	38 021	38 039	38 057	38 075	38 093	38 112	38 130	38 148	38 166	38 184
241	202	220	238	256	274	292	310	328	346	364
242	382	399	417	435	453	471	489	507	525	543
243	561	578	596	614	632	650	668	686	703	721
244	739	757	775	792	810	828	846	863	881	899
245	917	934	952	970	987	39 005	39 023	39 041	39 058	39 076
246	39 094	39 111	39 129	39 146	39 164	182	199	217	235	252
247	270	287	305	322	340	358	375	393	410	428
248	445	463	480	498	515	533	550	568	585	602
249	620	637	655	672	690	707	724	742	759	777
250	39 794	39 811	39 829	39 846	39 863	39 881	39 898	39 915	39 933	39 950
N	0	1	2	3	4	5	6	7	8	9

P.P.

	22	21
1	2.2	2.1
2	4.4	4.2
3	6.6	6.3
4	8.8	8.4
5	11.0	10.5
6	13.2	12.6
7	15.4	14.7
8	17.6	16.8
9	19.8	18.9

	20
1	2.0
2	4.0
3	6.0
4	8.0
5	10.0
6	12.0
7	14.0
8	16.0
9	18.0

	19
1	1.9
2	3.8
3	5.7
4	7.6
5	9.5
6	11.4
7	13.3
8	15.2
9	17.1

	18
1	1.8
2	3.6
3	5.4
4	7.2
5	9.0
6	10.8
7	12.6
8	14.4
9	16.2

	17
1	1.7
2	3.4
3	5.1
4	6.8
5	8.5
6	10.2
7	11.9
8	13.6
9	15.3

Numbers 250–300 Logs 39794–47842

N	0	1	2	3	4	5	6	7	8	9
250	39 794	39 811	39 829	39 846	39 863	39 881	39 898	39 915	39 933	39 950
251	967	985	40 002	40 019	40 037	40 054	40 071	40 088	40 106	40 123
252	40 140	40 157	175	192	209	226	243	261	278	295
253	312	329	346	364	381	398	415	432	449	466
254	483	500	518	535	552	569	586	603	620	637
255	654	671	688	705	722	739	756	773	790	807
256	824	841	858	875	892	909	926	943	960	976
257	993	41 010	41 027	41 044	41 061	41 078	41 095	41 111	41 128	41 145
258	41 162	179	196	212	229	246	263	280	296	313
259	330	347	363	380	397	414	430	447	464	481
260	41 497	41 514	41 531	41 547	41 564	41 581	41 597	41 614	41 631	41 647
261	664	681	697	714	731	747	764	780	797	814
262	830	847	863	880	896	913	929	946	963	979
263	996	42 012	42 029	42 045	42 062	42 078	42 095	42 111	42 127	42 144
264	42 160	177	193	210	226	243	259	275	292	308
265	325	341	357	374	390	406	423	439	455	472
266	488	504	521	537	553	570	586	602	619	635
267	651	667	684	700	716	732	749	765	781	797
268	813	830	846	862	878	894	911	927	943	959
269	975	991	43 008	43 024	43 040	43 056	43 072	43 088	43 104	43 120
270	43 136	43 152	43 169	43 185	43 201	43 217	43 233	43 249	43 265	43 281
271	297	313	329	345	361	377	393	409	425	441
272	457	473	489	505	521	537	553	569	584	600
273	616	632	648	664	680	696	712	727	743	759
274	775	791	807	823	838	854	870	886	902	917
275	933	949	965	981	996	44 012	44 028	44 044	44 059	44 075
276	44 091	44 107	44 122	44 138	44 154	170	185	201	217	232
277	248	264	279	295	311	326	342	358	373	389
278	404	420	436	451	467	483	498	514	529	545
279	560	576	592	607	623	638	654	669	685	700
280	44 716	44 731	44 747	44 762	44 778	44 793	44 809	44 824	44 840	44 855
281	871	886	902	917	932	948	963	979	994	45 010
282	45 025	45 040	45 056	45 071	45 086	45 102	45 117	45 133	45 148	163
283	179	194	209	225	240	255	271	286	301	317
284	332	347	362	378	393	408	423	439	454	469
285	484	500	515	530	545	561	576	591	606	621
286	637	652	667	682	697	712	728	743	758	773
287	788	803	818	834	849	864	879	894	909	924
288	939	954	969	984	46 000	46 015	46 030	46 045	46 060	46 075
289	46 090	46 105	46 120	46 135	150	165	180	195	210	225
290	46 240	46 255	46 270	46 285	46 300	46 315	46 330	46 345	46 359	46 374
291	389	404	419	434	449	464	479	494	509	523
292	538	553	568	583	598	613	627	642	657	672
293	687	702	716	731	746	761	776	790	805	820
294	835	850	864	879	894	909	923	938	953	967
295	982	997	47 012	47 026	47 041	47 056	47 070	47 085	47 100	47 114
296	47 129	47 144	159	173	188	202	217	232	246	261
297	276	290	305	319	334	349	363	378	392	407
298	422	436	451	465	480	494	509	524	538	553
299	567	582	596	611	625	640	654	669	683	698
300	47 712	47 727	47 741	47 756	47 770	47 784	47 799	47 813	47 828	47 842
N	0	1	2	3	4	5	6	7	8	9

P.P.

	18	17	16	15	14
1	1.8	1.7	1.6	1.5	1.4
2	3.6	3.4	3.2	3.0	2.8
3	5.4	5.1	4.8	4.5	4.2
4	7.2	6.8	6.4	6.0	5.6
5	9.0	8.5	8.0	7.5	7.0
6	10.8	10.2	9.6	9.0	8.4
7	12.6	11.9	11.2	10.5	9.8
8	14.4	13.6	12.8	12.0	11.2
9	16.2	15.3	14.4	13.5	12.6

Numbers 300–350 Logs 47712–54518

N	0	1	2	3	4	5	6	7	8	9
300	47 712	47 727	47 741	47 756	47 770	47 784	47 799	47 813	47 828	47 842
301	857	871	885	900	914	929	943	958	972	986
302	48 001	48 015	48 029	48 044	48 058	48 073	48 087	48 101	48 116	48 130
303	144	159	173	187	202	216	230	244	259	273
304	287	302	316	330	344	359	373	387	401	416
305	430	444	458	473	487	501	515	530	544	558
306	572	586	601	615	629	643	657	671	686	700
307	714	728	742	756	770	785	799	813	827	841
308	855	869	883	897	911	926	940	954	968	982
309	996	49 010	49 024	49 038	49 052	49 066	49 080	49 094	49 108	49 122
310	49 136	49 150	49 164	49 178	49 192	49 206	49 220	49 234	49 248	49 262
311	276	290	304	318	332	346	360	374	388	402
312	415	429	443	457	471	485	499	513	527	541
313	554	568	582	596	610	624	638	651	665	679
314	693	707	721	734	748	762	776	790	803	817
315	831	845	859	872	886	900	914	927	941	955
316	969	982	996	50 010	50 024	50 037	50 051	50 065	50 079	50 092
317	50 106	50 120	50 133	147	161	174	188	202	215	229
318	243	256	270	284	297	311	325	338	352	365
319	379	393	406	420	433	447	461	474	488	501
320	50 515	50 529	50 542	50 556	50 569	50 583	50 596	50 610	50 623	50 637
321	651	664	678	691	705	718	732	745	759	772
322	786	799	813	826	840	853	866	880	893	907
323	920	934	947	961	974	987	51 001	51 014	51 028	51 041
324	51 055	51 068	51 081	51 095	51 108	51 121	135	148	162	175
325	188	202	215	228	242	255	268	282	295	308
326	322	335	348	362	375	388	402	415	428	441
327	455	468	481	495	508	521	534	548	561	574
328	587	601	614	627	640	654	667	680	693	706
329	720	733	746	759	772	786	799	812	825	838
330	51 851	51 865	51 878	51 891	51 904	51 917	51 930	51 943	51 957	51 970
331	983	996	52 009	52 022	52 035	52 048	52 061	52 075	52 088	52 101
332	52 114	52 127	140	153	166	179	192	205	218	231
333	244	257	270	284	297	310	323	336	349	362
334	375	388	401	414	427	440	453	466	479	492
335	504	517	530	543	556	569	582	595	608	621
336	634	647	660	673	686	699	711	724	737	750
337	763	776	789	802	815	827	840	853	866	879
338	892	905	917	930	943	956	969	982	994	53 007
339	53 020	53 033	53 046	53 058	53 071	53 084	53 097	53 110	53 122	135
340	53 148	53 161	53 173	53 186	53 199	53 212	53 224	53 237	53 250	53 263
341	275	288	301	314	326	339	352	364	377	390
342	403	415	428	441	453	466	479	491	504	517
343	529	542	555	567	580	593	605	618	631	643
344	656	668	681	694	706	719	732	744	757	769
345	782	794	807	820	832	845	857	870	882	895
346	908	920	933	945	958	970	983	995	54 008	54 020
347	54 033	54 045	54 058	54 070	54 083	54 095	54 108	54 120	133	145
348	158	170	183	195	208	220	233	245	258	270
349	283	295	307	320	332	345	357	370	382	394
350	54 407	54 419	54 432	54 444	54 456	54 469	54 481	54 494	54 506	54 518
N	0	1	2	3	4	5	6	7	8	9

P.P.

	15	14	13	12
1	1.5	1.4	1.3	1.2
2	3.0	2.8	2.6	2.4
3	4.5	4.2	3.9	3.6
4	6.0	5.6	5.2	4.8
5	7.5	7.0	6.5	6.0
6	9.0	8.4	7.8	7.2
7	10.5	9.8	9.1	8.4
8	12.0	11.2	10.4	9.6
9	13.5	12.6	11.7	10.8

Numbers 350–400 Logs 54407–60304

N	0	1	2	3	4	5	6	7	8	9
350	54 407	54 419	54 432	54 444	54 456	54 469	54 481	54 494	54 506	54 518
351	531	543	555	568	580	593	605	617	630	642
352	654	667	679	691	704	716	728	741	753	765
353	777	790	802	814	827	839	851	864	876	888
354	900	913	925	937	949	962	974	986	998	55 011
355	55 023	55 035	55 047	55 060	55 072	55 084	55 096	55 108	55 121	133
356	145	157	169	182	194	206	218	230	242	255
357	267	279	291	303	315	328	340	352	364	376
358	388	400	413	425	437	449	461	473	485	497
359	509	522	534	546	558	570	582	594	606	618
360	55 630	55 642	55 654	55 666	55 678	55 691	55 703	55 715	55 727	55 739
361	751	763	775	787	799	811	823	835	847	859
362	871	883	895	907	919	931	943	955	967	979
363	991	56 003	56 015	56 027	56 038	56 050	56 062	56 074	56 086	56 098
364	56 110	122	134	146	158	170	182	194	205	217
365	229	241	253	265	277	289	301	312	324	336
366	348	360	372	384	396	407	419	431	443	455
367	467	478	490	502	514	526	538	549	561	573
368	585	597	608	620	632	644	656	667	679	691
369	703	714	726	738	750	761	773	785	797	808
370	56 820	56 832	56 844	56 855	56 867	56 879	56 891	56 902	56 914	56 926
371	937	949	961	972	984	996	57 008	57 019	57 031	57 043
372	57 054	57 066	57 078	57 089	57 101	57 113	124	136	148	159
373	171	183	194	206	217	229	241	252	264	276
374	287	299	310	322	334	345	357	368	380	392
375	403	415	426	438	449	461	473	484	496	507
376	519	530	542	553	565	576	588	600	611	623
377	634	646	657	669	680	692	703	715	726	738
378	749	761	772	784	795	807	818	830	841	852
379	864	875	887	898	910	921	933	944	955	967
380	57 978	57 990	58 001	58 013	58 024	58 035	58 047	58 058	58 070	58 081
381	58 092	58 104	115	127	138	149	161	172	184	195
382	206	218	229	240	252	263	274	286	297	309
383	320	331	343	354	365	377	388	399	410	422
384	433	444	456	467	478	490	501	512	524	535
385	546	557	569	580	591	602	614	625	636	647
386	659	670	681	692	704	715	726	737	749	760
387	771	782	794	805	816	827	838	850	861	872
388	883	894	906	917	928	939	950	961	973	984
389	995	59 006	59 017	59 028	59 040	59 051	59 062	59 073	59 084	59 095
390	59 106	59 118	59 129	59 140	59 151	59 162	59 173	59 184	59 195	59 207
391	218	229	240	251	262	273	284	295	306	318
392	329	340	351	362	373	384	395	406	417	428
393	439	450	461	472	483	494	506	517	528	539
394	550	561	572	583	594	605	616	627	638	649
395	660	671	682	693	704	715	726	737	748	759
396	770	780	791	802	813	824	835	846	857	868
397	879	890	901	912	923	934	945	956	966	977
398	988	999	60 010	60 021	60 032	60 043	60 054	60 065	60 076	60 086
399	60 097	60 108	119	130	141	152	163	173	184	195
400	60 206	60 217	60 228	60 239	60 249	60 260	60 271	60 282	60 293	60 304
N	0	1	2	3	4	5	6	7	8	9

P.P.

13	
1	1.3
2	2.6
3	3.9
4	5.2
5	6.5
6	7.8
7	9.1
8	10.4
9	11.7

12	
1	1.2
2	2.4
3	3.6
4	4.8
5	6.0
6	7.2
7	8.4
8	9.6
9	10.8

11	
1	1.1
2	2.2
3	3.3
4	4.4
5	5.5
6	6.6
7	7.7
8	8.8
9	9.9

10	
1	1.0
2	2.0
3	3.0
4	4.0
5	5.0
6	6.0
7	7.0
8	8.0
9	9.0

Numbers 400–450 Logs 60206–65408

N	0	1	2	3	4	5	6	7	8	9	P.P.	
400	60 206	60 217	60 228	60 239	60 249	60 260	60 271	60 282	60 293	60 304		11
401	314	325	336	347	358	369	379	390	401	412	1	1.1
402	423	433	444	455	466	477	487	498	509	520	2	2.2
403	531	541	552	563	574	584	595	606	617	627	3	3.3
404	638	649	660	670	681	692	703	713	724	735	4	4.4
405	746	756	767	778	788	799	810	821	831	842	5	5.5
406	853	863	874	885	895	906	917	927	938	949	6	6.6
407	959	970	981	991	61 002	61 013	61 023	61 034	61 045	61 055	7	7.7
408	61 066	61 077	61 087	61 098	109	119	130	140	151	162	8	8.8
409	172	183	194	204	215	225	236	247	257	268	9	9.9
410	61 278	61 289	61 300	61 310	61 321	61 331	61 342	61 352	61 363	61 374		
411	384	395	405	416	426	437	448	458	469	479		
412	490	500	511	521	532	542	553	563	574	584		
413	595	606	616	627	637	648	658	669	679	690		
414	700	711	721	731	742	752	763	773	784	794		
415	805	815	826	836	847	857	868	878	888	899		
416	909	920	930	941	951	962	972	982	993	62 003		
417	62 014	62 024	62 034	62 045	62 055	62 066	62 076	62 086	62 097	107		
418	118	128	138	149	159	170	180	190	201	211		
419	221	232	242	252	263	273	284	294	304	315		
420	62 325	62 335	62 346	62 356	62 366	62 377	62 387	62 397	62 408	62 418		10
421	428	439	449	459	469	480	490	500	511	521	1	1.0
422	531	542	552	562	572	583	593	603	613	624	2	2.0
423	634	644	655	665	675	685	696	706	716	726	3	3.0
424	737	747	757	767	778	788	798	808	818	829	4	4.0
425	839	849	859	870	880	890	900	910	921	931	5	5.0
426	941	951	961	972	982	992	63 002	63 012	63 022	63 033	6	6.0
427	63 043	63 053	63 063	63 073	63 083	63 094	104	114	124	134	7	7.0
428	144	155	165	175	185	195	205	215	225	236	8	8.0
429	246	256	266	276	286	296	306	317	327	337	9	9.0
430	63 347	63 357	63 367	63 377	63 387	63 397	63 407	63 417	63 428	63 438		
431	448	458	468	478	488	498	508	518	528	538		
432	548	558	568	579	589	599	609	619	629	639		
433	649	659	669	679	689	699	709	719	729	739		
434	749	759	769	779	789	799	809	819	829	839		
435	849	859	869	879	889	899	909	919	929	939		
436	949	959	969	979	988	998	64 008	64 018	64 028	64 038		
437	64 048	64 058	64 068	64 078	64 088	64 098	108	118	128	137		
438	147	157	167	177	187	197	207	217	227	237		
439	246	256	266	276	286	296	306	316	326	335		
440	64 345	64 355	64 365	64 375	64 385	64 395	64 404	64 414	64 424	64 434		9
441	444	454	464	473	483	493	503	513	523	532	1	.9
442	542	552	562	572	582	591	601	611	621	631	2	1.8
443	640	650	660	670	680	689	699	709	719	729	3	2.7
444	738	748	758	768	777	787	797	807	816	826	4	3.6
445	836	846	856	865	875	885	895	904	914	924	5	4.5
446	933	943	953	963	972	982	992	65 002	65 011	65 021	6	5.4
447	65 031	65 040	65 050	65 060	65 070	65 079	65 089	099	108	118	7	6.3
448	128	137	147	157	167	176	186	196	205	215	8	7.2
449	225	234	244	254	263	273	283	292	302	312	9	8.1
450	65 321	65 331	65 341	65 350	65 360	65 369	65 379	65 389	65 398	65 408		
N	0	1	2	3	4	5	6	7	8	9		

Numbers 450–500 Logs 65321–69975

N	0	1	2	3	4	5	6	7	8	9	P.P.	
450	65 321	65 331	65 341	65 350	65 360	65 369	65 379	65 389	65 398	65 408	10	
451	418	427	437	447	456	466	475	485	495	504	1	1.0
452	514	523	533	543	552	562	571	581	591	600	2	2.0
453	610	619	629	639	648	658	667	677	686	696	3	3.0
454	706	715	725	734	744	753	763	772	782	792	4	4.0
455	801	811	820	830	839	849	858	868	877	887	5	5.0
456	896	906	916	925	935	944	954	963	973	982	6	6.0
457	992	66 001	66 011	66 020	66 030	66 039	66 049	66 058	66 068	66 077	7	7.0
458	66 087	096	106	115	124	134	143	153	162	172	8	8.0
459	181	191	200	210	219	229	238	247	257	266	9	9.0
460	66 276	66 285	66 295	66 304	66 314	66 323	66 332	66 342	66 351	66 361		
461	370	380	389	398	408	417	427	436	445	455		
462	464	474	483	492	502	511	521	530	539	549		
463	558	567	577	586	596	605	614	624	633	642		
464	652	661	671	680	689	699	708	717	727	736		
465	745	755	764	773	783	792	801	811	820	829		
466	839	848	857	867	876	885	894	904	913	922		
467	932	941	950	960	969	978	987	997	67 006	67 015		
468	67 025	67 034	67 043	67 052	67 062	67 071	67 080	67 089	099	108		
469	117	127	136	145	154	164	173	182	191	201		
470	67 210	67 219	67 228	67 237	67 247	67 256	67 265	67 274	67 284	67 293	9	
471	302	311	321	330	339	348	357	367	376	385	1	0.9
472	394	403	413	422	431	440	449	459	468	477	2	1.8
473	486	495	504	514	523	532	541	550	560	569	3	2.7
474	578	587	596	605	614	624	633	642	651	660	4	3.6
475	669	679	688	697	706	715	724	733	742	752	5	4.5
476	761	770	779	788	797	806	815	825	834	843	6	5.4
477	852	861	870	879	888	897	906	916	925	934	7	6.3
478	943	952	961	970	979	988	997	68 006	68 015	68 024	8	7.2
479	68 034	68 043	68 052	68 061	68 070	68 079	68 088	097	106	115	9	8.1
480	68 124	68 133	68 142	68 151	68 160	68 169	68 178	68 187	68 196	68 205		
481	215	224	233	242	251	260	269	278	287	296		
482	305	314	323	332	341	350	359	368	377	386		
483	395	404	413	422	431	440	449	458	467	476		
484	485	494	502	511	520	529	538	547	556	565		
485	574	583	592	601	610	619	628	637	646	655		
486	664	673	681	690	699	708	717	726	735	744		
487	753	762	771	780	789	797	806	815	824	833		
488	842	851	860	869	878	886	895	904	913	922		
489	931	940	949	958	966	975	984	993	69 002	69 011		
490	69 020	69 028	69 037	69 046	69 055	69 064	69 073	69 082	69 090	69 099	8	
491	108	117	126	135	144	152	161	170	179	188	1	0.8
492	197	205	214	223	232	241	249	258	267	276	2	1.6
493	285	294	302	311	320	329	338	346	355	364	3	2.4
494	373	381	390	399	408	417	425	434	443	452	4	3.2
495	461	469	478	487	496	504	513	522	531	539	5	4.0
496	548	557	566	574	583	592	601	609	618	627	6	4.8
497	636	644	653	662	671	679	688	697	705	714	7	5.6
498	723	732	740	749	758	767	775	784	793	801	8	6.4
499	810	819	827	836	845	854	862	871	880	888	9	7.2
500	69 897	69 906	69 914	69 923	69 932	69 940	69 949	69 958	69 966	69 975		
N	0	1	2	3	4	5	6	7	8	9		

Numbers 500–550 Logs 69897–74107

N	0	1	2	3	4	5	6	7	8	9	P.P.
500	69 897	69 906	69 914	69 923	69 932	69 940	69 949	69 958	69 966	69 975	9
501	984	992	70 001	70 010	70 018	70 027	70 036	70 044	70 053	70 062	1 0.9
502	70 070	70 079	088	096	105	114	122	131	140	148	2 1.8
503	157	165	174	183	191	200	209	217	226	234	3 2.7
504	243	252	260	269	278	286	295	303	312	321	4 3.6
505	329	338	346	355	364	372	381	389	398	406	5 4.5
506	415	424	432	441	449	458	467	475	484	492	6 5.4
507	501	509	518	526	535	544	552	561	569	578	7 6.3
508	586	595	603	612	621	629	638	646	655	663	8 7.2
509	672	680	689	697	706	714	723	731	740	749	9 8.1
510	70 757	70 766	70 774	70 783	70 791	70 800	70 808	70 817	70 825	70 834	
511	842	851	859	868	876	885	893	902	910	919	
512	927	935	944	952	961	969	978	986	995	71 003	
513	71 012	71 020	71 029	71 037	71 046	71 054	71 063	71 071	71 079	088	
514	096	105	113	122	130	139	147	155	164	172	
515	181	189	198	206	214	223	231	240	248	257	
516	265	273	282	290	299	307	315	324	332	341	
517	349	357	366	374	383	391	399	408	416	425	
518	433	441	450	458	466	475	483	492	500	508	
519	517	525	533	542	550	559	567	575	584	592	
520	71 600	71 609	71 617	71 625	71 634	71 642	71 650	71 659	71 667	71 675	8
521	684	692	700	709	717	725	734	742	750	759	1 0.8
522	767	775	784	792	800	809	817	825	834	842	2 1.6
523	850	858	867	875	883	892	900	908	917	925	3 2.4
524	933	941	950	958	966	975	983	991	999	72 008	4 3.2
525	72 016	72 024	72 032	72 041	72 049	72 057	72 066	72 074	72 082	090	5 4.0
526	099	107	115	123	132	140	148	156	165	173	6 4.8
527	181	189	198	206	214	222	230	239	247	255	7 5.6
528	263	272	280	288	296	304	313	321	329	337	8 6.4
529	346	354	362	370	378	387	395	403	411	419	9 7.2
530	72 428	72 436	72 444	72 452	72 460	72 469	72 477	72 485	72 493	72 501	
531	509	518	526	534	542	550	558	567	575	583	
532	591	599	607	616	624	632	640	648	656	665	
533	673	681	689	697	705	713	722	730	738	746	
534	754	762	770	779	787	795	803	811	819	827	
535	835	843	852	860	868	876	884	892	900	908	
536	916	925	933	941	949	957	965	973	981	989	
537	997	73 006	73 014	73 022	73 030	73 038	73 046	73 054	73 062	73 070	
538	73 078	086	094	102	111	119	127	135	143	151	
539	159	167	175	183	191	199	207	215	223	231	
540	73 239	73 247	73 255	73 263	73 272	73 280	73 288	73 296	73 304	73 312	7
541	320	328	336	344	352	360	368	376	384	392	1 0.7
542	400	408	416	424	432	440	448	456	464	472	2 1.4
543	480	488	496	504	512	520	528	536	544	552	3 2.1
544	560	568	576	584	592	600	608	616	624	632	4 2.8
545	640	648	656	664	672	679	687	695	703	711	5 3.5
546	719	727	735	743	751	759	767	775	783	791	6 4.2
547	799	807	815	823	830	838	846	854	862	870	7 4.9
548	878	886	894	902	910	918	926	933	941	949	8 5.6
549	957	965	973	981	989	997	74 005	74 013	74 020	74 028	9 6.3
550	74 036	74 044	74 052	74 060	74 068	74 076	74 084	74 092	74 099	74 107	
N	0	1	2	3	4	5	6	7	8	9	

Numbers 550–600 Logs 74036–77880

N	0	1	2	3	4	5	6	7	8	9	P.P.
550	74 036	74 044	74 052	74 060	74 068	74 076	74 084	74 092	74 099	74 107	8
551	115	123	131	139	147	155	162	170	178	186	1 0.8
552	194	202	210	218	225	233	241	249	257	265	2 1.6
553	273	280	288	296	304	312	320	327	335	343	3 2.4
554	351	359	367	374	382	390	398	406	414	421	4 3.2
555	429	437	445	453	461	468	476	484	492	500	5 4.0
556	507	515	523	531	539	547	554	562	570	578	6 4.8
557	586	593	601	609	617	624	632	640	648	656	7 5.6
558	663	671	679	687	695	702	710	718	726	733	8 6.4
559	741	749	757	764	772	780	788	796	803	811	9 7.2
560	74 819	74 827	74 834	74 842	74 850	74 858	74 865	74 873	74 881	74 889	
561	896	904	912	920	927	935	943	950	958	966	
562	974	981	989	997	75 005	75 012	75 020	75 028	75 035	75 043	
563	75 051	75 059	75 066	75 074	082	089	097	105	113	120	
564	128	136	143	151	159	166	174	182	189	197	
565	205	213	220	228	236	243	251	259	266	274	
566	282	289	297	305	312	320	328	335	343	351	
567	358	366	374	381	389	397	404	412	420	427	
568	435	442	450	458	465	473	481	488	496	504	
569	511	519	526	534	542	549	557	565	572	580	
570	75 587	75 595	75 603	75 610	75 618	75 626	75 633	75 641	75 648	75 656	7
571	664	671	679	686	694	702	709	717	724	732	1 0.7
572	740	747	755	762	770	778	785	793	800	808	2 1.4
573	815	823	831	838	846	853	861	868	876	884	3 2.1
574	891	899	906	914	921	929	937	944	952	959	4 2.8
575	967	974	982	989	997	76005	76 012	76 020	76 027	76 035	5 3.5
576	76 042	76 050	76 057	76 065	76 072	080	087	095	103	110	6 4.2
577	118	125	133	140	148	155	163	170	178	185	7 4.9
578	193	200	208	215	223	230	238	245	253	260	8 5.6
579	268	275	283	290	298	305	313	320	328	335	9 6.3
580	76 343	76 350	76 358	76 365	76 373	76 380	76 388	76 395	76 403	76 410	
581	418	425	433	440	448	455	462	470	477	485	
582	492	500	507	515	522	530	537	545	552	559	
583	567	574	582	589	597	604	612	619	626	634	
584	641	649	656	664	671	678	686	693	701	708	
585	716	723	730	738	745	753	760	768	775	782	
586	790	797	805	812	819	827	834	842	849	856	
587	864	871	879	886	893	901	908	916	923	930	
588	938	945	953	960	967	975	982	989	997	77 004	
589	77 012	77 019	77 026	77 034	77 041	77 048	77 056	77 063	77 070	078	
590	77 085	77 093	77 100	77 107	77 115	77 122	77 129	77 137	77 144	77 151	
591	159	166	173	181	188	195	203	210	217	225	
592	232	240	247	254	262	269	276	283	291	298	
593	305	313	320	327	335	342	349	357	364	371	
594	379	386	393	401	408	415	422	430	437	444	
595	452	459	466	474	481	488	495	503	510	517	
596	525	532	539	546	554	561	568	576	583	590	
597	597	605	612	619	627	634	641	648	656	663	
598	670	677	685	692	699	706	714	721	728	735	
599	743	750	757	764	772	779	786	793	801	808	
600	77 815	77 822	77 830	77 837	77 844	77 851	77 859	77 866	77 873	77 880	
N	**0**	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	

Numbers 600–650 Logs 77815–81351

N	0	1	2	3	4	5	6	7	8	9	P.P.
600	77 815	77 822	77 830	77 837	77 844	77 851	77 859	77 866	77 873	77 880	8
601	887	895	902	909	916	924	931	938	945	952	1 0.8
602	960	967	974	981	988	996	78 003	78 010	78 017	78 025	2 1.6
603	78 032	78 039	78 046	78 053	78 061	78 068	075	082	089	097	3 2.4
604	104	111	118	125	132	140	147	154	161	168	4 3.2
605	176	183	190	197	204	211	219	226	233	240	5 4.0
606	247	254	262	269	276	283	290	297	305	312	6 4.8
607	319	326	333	340	347	355	362	369	376	383	7 5.6
608	390	398	405	412	419	426	433	440	447	455	8 6.4
609	462	469	476	483	490	497	504	512	519	526	9 7.2
610	78 533	78 540	78 547	78 554	78 561	78 569	78 576	78 583	78 590	78 597	
611	604	611	618	625	633	640	647	654	661	668	
612	675	682	689	696	704	711	718	725	732	739	
613	746	753	760	767	774	781	789	796	803	810	
614	817	824	831	838	845	852	859	866	873	880	
615	888	895	902	909	916	923	930	937	944	951	
616	958	965	972	979	986	993	79 000	79 007	79 014	79 021	
617	79 029	79 036	79 043	79 050	79 057	79 064	071	078	085	092	
618	099	106	113	120	127	134	141	148	155	162	
619	169	176	183	190	197	204	211	218	225	232	
620	79 239	79 246	79 253	79 260	79 267	79 274	79 281	79 288	79 295	79 302	7
621	309	316	323	330	337	344	351	358	365	372	1 0.7
622	379	386	393	400	407	414	421	428	435	442	2 1.4
623	449	456	463	470	477	484	491	498	505	511	3 2.1
624	518	525	532	539	546	553	560	567	574	581	4 2.8
625	588	595	602	609	616	623	630	637	644	650	5 3.5
626	657	664	671	678	685	692	699	706	713	720	6 4.2
627	727	734	741	748	754	761	768	775	782	789	7 4.9
628	796	803	810	817	824	831	837	844	851	858	8 5.6
629	865	872	879	886	893	900	906	913	920	927	9 6.3
630	79 934	79 941	79 948	79 955	79 962	79 969	79 975	79 982	79 989	79 996	
631	80 003	80 010	80 017	80 024	80 030	80 037	80 044	80 051	80 058	80 065	
632	072	079	085	092	099	106	113	120	127	134	
633	140	147	154	161	168	175	182	188	195	202	
634	209	216	223	229	236	243	250	257	264	271	
635	277	284	291	298	305	312	318	325	332	339	
636	346	353	359	366	373	380	387	393	400	407	
637	414	421	428	434	441	448	455	462	468	475	
638	482	489	496	502	509	516	523	530	536	543	
639	550	557	564	570	577	584	591	598	604	611	
640	80 618	80 625	80 632	80 638	80 645	80 652	80 659	80 665	80 672	80 679	6
641	686	693	699	706	713	720	726	733	740	747	1 0.6
642	754	760	767	774	781	787	794	801	808	814	2 1.2
643	821	828	835	841	848	855	862	868	875	882	3 1.8
644	889	895	902	909	916	922	929	936	943	949	4 2.4
645	956	963	969	976	983	990	996	81 003	81 010	81 017	5 3.0
646	81 023	81 030	81 037	81 043	81 050	81 057	81 064	070	077	084	6 3.6
647	090	097	104	111	117	124	131	137	144	151	7 4.2
648	158	164	171	178	184	191	198	204	211	218	8 4.8
649	224	231	238	245	251	258	265	271	278	285	9 5.4
650	81 291	81 298	81 305	81 311	81 318	81 325	81 331	81 338	81 345	81 351	
N	0	1	2	3	4	5	6	7	8	9	

Numbers 650–700 Logs 81291–84566

N	0	1	2	3	4	5	6	7	8	9	P.P.
650	81 291	81 298	81 305	81 311	81 318	81 325	81 331	81 338	81 345	81 351	7
651	358	365	371	378	385	391	398	405	411	418	1 0.7
652	425	431	438	445	451	458	465	471	478	485	2 1.4
653	491	498	505	511	518	525	531	538	544	551	3 2.1
654	558	564	571	578	584	591	598	604	611	617	4 2.8
655	624	631	637	644	651	657	664	671	677	684	5 3.5
656	690	697	704	710	717	723	730	737	743	750	6 4.2
657	757	763	770	776	783	790	796	803	809	816	7 4.9
658	823	829	836	842	849	856	862	869	875	882	8 5.6
659	889	895	902	908	915	921	928	935	941	948	9 6.3
660	81 954	81 961	81 968	81 974	81 981	81 987	81 994	82 000	82 007	82 014	
661	82 020	82 027	82 033	82 040	82 046	82 053	82 060	066	073	079	
662	086	092	099	105	112	119	125	132	138	145	
663	151	158	164	171	178	184	191	197	204	210	
664	217	223	230	236	243	249	256	263	269	276	
665	282	289	295	302	308	315	321	328	334	341	
666	347	354	360	367	373	380	387	393	400	406	
667	413	419	426	432	439	445	452	458	465	471	
668	478	484	491	497	504	510	517	523	530	536	
669	543	549	556	562	569	575	582	588	595	601	
670	82 607	82 614	82 620	82 627	82 633	82 640	82 646	82 653	82 659	82 666	6
671	672	679	685	692	698	705	711	718	724	730	1 0.6
672	737	743	750	756	763	769	776	782	789	795	2 1.2
673	802	808	814	821	827	834	840	847	853	860	3 1.8
674	866	872	879	885	892	898	905	911	918	924	4 2.4
675	930	937	943	950	956	963	969	975	982	988	5 3.0
676	995	83 001	83 008	83 014	83 020	83 027	83 033	83 040	83 046	83 052	6 3.6
677	83 059	065	072	078	085	091	097	104	110	117	7 4.2
678	123	129	136	142	149	155	161	168	174	181	8 4.8
679	187	193	200	206	213	219	225	232	238	245	9 5.4
680	83 251	83 257	83 264	83 270	83 276	83 283	83 289	83 296	83 302	83 308	
681	315	321	327	334	340	347	353	359	366	372	
682	378	385	391	398	404	410	417	423	429	436	
683	442	448	455	461	467	474	480	487	493	499	
684	506	512	518	525	531	537	544	550	556	563	
685	569	575	582	588	594	601	607	613	620	626	
686	632	639	645	651	658	664	670	677	683	689	
687	696	702	708	715	721	727	734	740	746	753	
688	759	765	771	778	784	790	797	803	809	816	
689	822	828	835	841	847	853	860	866	872	879	
690	83 885	83 891	83 897	83 904	83 910	83 916	83 923	83 929	83 935	83 942	
691	948	954	960	967	973	979	985	992	998	84 004	
692	84 011	84 017	84 023	84 029	84 036	84 042	84 048	84 055	84 061	067	
693	073	080	086	092	098	105	111	117	123	130	
694	136	142	148	155	161	167	173	180	186	192	
695	198	205	211	217	223	230	236	242	248	255	
696	261	267	273	280	286	292	298	305	311	317	
697	323	330	336	342	348	354	361	367	373	379	
698	386	392	398	404	410	417	423	429	435	442	
699	448	454	460	466	473	479	485	491	497	504	
700	84 510	84 516	84 522	84 528	84 535	84 541	84 547	84 553	84 559	84 566	
N	0	1	2	3	4	5	6	7	8	9	

Numbers 700–750 Logs 84510–87558

N	0	1	2	3	4	5	6	7	8	9	P.P.
700	84 510	84 516	84 522	84 528	84 535	84 541	84 547	84 553	84 559	84 566	7
701	572	578	584	590	597	603	609	615	621	628	1 0.7
702	634	640	646	652	658	665	671	677	683	689	2 1.4
703	696	702	708	714	720	726	733	739	745	751	3 2.1
704	757	763	770	776	782	788	794	800	807	813	4 2.8
705	819	825	831	837	844	850	856	862	868	874	5 3.5
706	880	887	893	899	905	911	917	924	930	936	6 4.2
707	942	948	954	960	967	973	979	985	991	997	7 4.9
708	85 003	85 009	85 016	85 022	85 028	85 034	85 040	85 046	85 052	85 058	8 5.6
709	065	071	077	083	089	095	101	107	114	120	9 6.3
710	85 126	85 132	85 138	85 144	85 150	85 156	85 163	85 169	85 175	85 181	
711	187	193	199	205	211	217	224	230	236	242	
712	248	254	260	266	272	278	285	291	297	303	
713	309	315	321	327	333	339	345	352	358	364	
714	370	376	382	388	394	400	406	412	418	425	
715	431	437	443	449	455	461	467	473	479	485	
716	491	497	503	509	516	522	528	534	540	546	
717	552	558	564	570	576	582	588	594	600	606	
718	612	618	625	631	637	643	649	655	661	667	
719	673	679	685	691	697	703	709	715	721	727	
720	85 733	85 739	85 745	85 751	85 757	85 763	85 769	85 775	85 781	85 788	6
721	794	800	806	812	818	824	830	836	842	848	1 0.6
722	854	860	866	872	878	884	890	896	902	908	2 1.2
723	914	920	926	932	938	944	950	956	962	968	3 1.8
724	974	980	986	992	998	86 004	86 010	86 016	86 022	86 028	4 2.4
725	86 034	86 040	86 046	86 052	86 058	064	070	076	082	088	5 3.0
726	094	100	106	112	118	124	130	136	141	147	6 3.6
727	153	159	165	171	177	183	189	195	201	207	7 4.2
728	213	219	225	231	237	243	249	255	261	267	8 4.8
729	273	279	285	291	297	303	308	314	320	326	9 5.4
730	86 332	86 338	86 344	86 350	86 356	86 362	86 368	86 374	86 380	86 386	
731	392	398	404	410	415	421	427	433	439	445	
732	451	457	463	469	475	481	487	493	499	504	
733	510	516	522	528	534	540	546	552	558	564	
734	570	576	581	587	593	599	605	611	617	623	
735	629	635	641	646	652	658	664	670	676	682	
736	688	694	700	705	711	717	723	729	735	741	
737	747	753	759	764	770	776	782	788	794	800	
738	806	812	817	823	829	835	841	847	853	859	
739	864	870	876	882	888	894	900	906	911	917	
740	86 923	86 929	86 935	86 941	86 947	86 953	86 958	86 964	86 970	86 976	5
741	982	988	994	999	87 005	87 011	87 017	87 023	87 029	87 035	1 0.5
742	87 040	87 046	87 052	87 058	064	070	075	081	087	093	2 1.0
743	099	105	111	116	122	128	134	140	146	151	3 1.5
744	157	163	169	175	181	186	192	198	204	210	4 2.0
745	216	221	227	233	239	245	251	256	262	268	5 2.5
746	274	280	286	291	297	303	309	315	320	326	6 3.0
747	332	338	344	349	355	361	367	373	379	384	7 3.5
748	390	396	402	408	413	419	425	431	437	442	8 4.0
749	448	454	460	466	471	477	483	489	495	500	9 4.5
750	87 506	87 512	87 518	87 523	87 529	87 535	87 541	87 547	87 552	87 558	
N	**0**	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	

Numbers 750–800 Logs 87506–90358

N	0	1	2	3	4	5	6	7	8	9
750	87 506	87 512	87 518	87 523	87 529	87 535	87 541	87 547	87 552	87 558
751	564	570	576	581	587	593	599	604	610	616
752	622	628	633	639	645	651	656	662	668	674
753	679	685	691	697	703	708	714	720	726	731
754	737	743	749	754	760	766	772	777	783	789
755	795	800	806	812	818	823	829	835	841	846
756	852	858	864	869	875	881	887	892	898	904
757	910	915	921	927	933	938	944	950	955	961
758	967	973	978	984	990	996	88 001	88 007	88 013	88 018
759	88 024	88 030	88 036	88 041	88 047	88 053	058	064	070	076
760	88 081	88 087	88 093	88 098	88 104	88 110	88 116	88 121	88 127	88 133
761	138	144	150	156	161	167	173	178	184	190
762	195	201	207	213	218	224	230	235	241	247
763	252	258	264	270	275	281	287	292	298	304
764	309	315	321	326	332	338	343	349	355	360
765	366	372	377	383	389	395	400	406	412	417
766	423	429	434	440	446	451	457	463	468	474
767	480	485	491	497	502	508	513	519	525	530
768	536	542	547	553	559	564	570	576	581	587
769	593	598	604	610	615	621	627	632	638	643
770	88 649	88 655	88 660	88 666	88 672	88 677	88 683	88 689	88 694	88 700
771	705	711	717	722	728	734	739	745	750	756
772	762	767	773	779	784	790	795	801	807	812
773	818	824	829	835	840	846	852	857	863	868
774	874	880	885	891	897	902	908	913	919	925
775	930	936	941	947	953	958	964	969	975	981
776	986	992	997	89 003	89 009	89 014	89 020	89 025	89 031	89 037
777	89 042	89 048	89 053	059	064	070	076	081	087	092
778	098	104	109	115	120	126	131	137	143	148
779	154	159	165	170	176	182	187	193	198	204
780	89 209	89 215	89 221	89 226	89 232	89 237	89 243	89 248	89 254	89 260
781	265	271	276	282	287	293	298	304	310	315
782	321	326	332	337	343	348	354	360	365	371
783	376	382	387	393	398	404	409	415	421	426
784	432	437	443	448	454	459	465	470	476	481
785	487	492	498	504	509	515	520	526	531	537
786	542	548	553	559	564	570	575	581	586	592
787	597	603	609	614	620	625	631	636	642	647
788	653	658	664	669	675	680	686	691	697	702
789	708	713	719	724	730	735	741	746	752	757
790	89 763	89 768	89 774	89 779	89 785	89 790	89 796	89 801	89 807	89 812
791	818	823	829	834	840	845	851	856	862	867
792	873	878	883	889	894	900	905	911	916	922
793	927	933	938	944	949	955	960	966	971	977
794	982	988	993	998	90 004	90 009	90 015	90 020	90 026	90 031
795	90 037	90 042	90 048	90 053	059	064	069	075	080	086
796	091	097	102	108	113	119	124	129	135	140
797	146	151	157	162	168	173	179	184	189	195
798	200	206	211	217	222	227	233	238	244	249
799	255	260	266	271	276	282	287	293	298	304
800	90 309	90 314	90 320	90 325	90 331	90 336	90 342	90 347	90 352	90 358
N	0	1	2	3	4	5	6	7	8	9

P.P.

	6
1	0.6
2	1.2
3	1.8
4	2.4
5	3.0
6	3.6
7	4.2
8	4.8
9	5.4

	5
1	0.5
2	1.0
3	1.5
4	2.0
5	2.5
6	3.0
7	3.5
8	4.0
9	4.5

Numbers 800–850 Logs 90309–92988

N	0	1	2	3	4	5	6	7	8	9
800	90 309	90 314	90 320	90 325	90 331	90 336	90 342	90 347	90 352	90 358
801	363	369	374	380	385	390	396	401	407	412
802	417	423	428	434	439	445	450	455	461	466
803	472	477	482	488	493	499	504	509	515	520
804	526	531	536	542	547	553	558	563	569	574
805	580	585	590	596	601	607	612	617	623	628
806	634	639	644	650	655	660	666	671	677	682
807	687	693	698	703	709	714	720	725	730	736
808	741	747	752	757	763	768	773	779	784	789
809	795	800	806	811	816	822	827	832	838	843
810	90 849	90 854	90 859	90 865	90 870	90 875	90 881	90 886	90 891	90 897
811	902	907	913	918	924	929	934	940	945	950
812	956	961	966	972	977	982	988	993	998	91 004
813	91 009	91 014	91 020	91 025	91 030	91 036	91 041	91 046	91 052	057
814	062	068	073	078	084	089	094	100	105	110
815	116	121	126	132	137	142	148	153	158	164
816	169	174	180	185	190	196	201	206	212	217
817	222	228	233	238	243	249	254	259	265	270
818	275	281	286	291	297	302	307	312	318	323
819	328	334	339	344	350	355	360	365	371	376
820	91 381	91 387	91 392	91 397	91 403	91 408	91 413	91 418	91 424	91 429
821	434	440	445	450	455	461	466	471	477	482
822	487	492	498	503	508	514	519	524	529	535
823	540	545	551	556	561	566	572	577	582	587
824	593	598	603	609	614	619	624	630	635	640
825	645	651	656	661	666	672	677	682	687	693
826	698	703	709	714	719	724	730	735	740	745
827	751	756	761	766	772	777	782	787	793	798
828	803	808	814	819	824	829	834	840	845	850
829	855	861	866	871	876	882	887	892	897	903
830	91 908	91 913	91 918	91 924	91 929	91 934	91 939	91 944	91 950	91 955
831	960	965	971	976	981	986	991	997	92 002	92 007
832	92 012	92 018	92 023	92 028	92 033	92 038	92 044	92 049	054	059
833	065	070	075	080	085	091	096	101	106	111
834	117	122	127	132	137	143	148	153	158	163
835	169	174	179	184	189	195	200	205	210	215
836	221	226	231	236	241	247	252	257	262	267
837	273	278	283	288	293	298	304	309	314	319
838	324	330	335	340	345	350	355	361	366	371
839	376	381	387	392	397	402	407	412	418	423
840	92 428	92 433	92 438	92 443	92 449	92 454	92 459	92 464	92 469	92 474
841	480	485	490	495	500	505	511	516	521	526
842	531	536	542	547	552	557	562	567	572	578
843	583	588	593	598	603	609	614	619	624	629
844	634	639	645	650	655	660	665	670	675	681
845	686	691	696	701	706	711	716	722	727	732
846	737	742	747	752	758	763	768	773	778	783
847	788	793	799	804	809	814	819	824	829	834
848	840	845	850	855	860	865	870	875	881	886
849	891	896	901	906	911	916	921	927	932	937
850	92 942	92 947	92 952	92 957	92 962	92 967	92 973	92 978	92 983	92 988
N	0	1	2	3	4	5	6	7	8	9

P.P.

	6
1	0.6
2	1.2
3	1.8
4	2.4
5	3.0
6	3.6
7	4.2
8	4.8
9	5.4

	5
1	0.5
2	1.0
3	1.5
4	2.0
5	2.5
6	3.0
7	3.5
8	4.0
9	4.5

Numbers 850–900 Logs 92942–95468

N	0	1	2	3	4	5	6	7	8	9	P.P.
850	92 942	92 947	92 952	92 957	92 962	92 967	92 973	92 978	92 983	92 988	6
851	993	998	93 003	93 008	93 013	93 018	93 024	93 029	93 034	93 039	1 0.6
852	93 044	93 049	054	059	064	069	075	080	085	090	2 1.2
853	095	100	105	110	115	120	125	131	136	141	3 1.8
854	146	151	156	161	166	171	176	181	186	192	4 2.4
855	197	202	207	212	217	222	227	232	237	242	5 3.0
856	247	252	258	263	268	273	278	283	288	293	6 3.6
857	298	303	308	313	318	323	328	334	339	344	7 4.2
858	349	354	359	364	369	374	379	384	389	394	8 4.8
859	399	404	409	414	420	425	430	435	440	445	9 5.4
860	93 450	93 455	93 460	93 465	93 470	93 475	93 480	93 485	93 490	93 495	
861	500	505	510	515	520	526	531	536	541	546	
862	551	556	561	566	571	576	581	586	591	596	
863	601	606	611	616	621	626	631	636	641	646	
864	651	656	661	666	671	676	682	687	692	697	
865	702	707	712	717	722	727	732	737	742	747	
866	752	757	762	767	772	777	782	787	792	797	
867	802	807	812	817	822	827	832	837	842	847	
868	852	857	862	867	872	877	882	887	892	897	
869	902	907	912	917	922	927	932	937	942	947	
870	93 952	93 957	93 962	93 967	93 972	93 977	93 982	93 987	93 992	93 997	5
871	94 002	94 007	94 012	94 017	94 022	94 027	94 032	94 037	94 042	94 047	1 0.5
872	052	057	062	067	072	077	082	086	091	096	2 1.0
873	101	106	111	116	121	126	131	136	141	146	3 1.5
874	151	156	161	166	171	176	181	186	191	196	4 2.0
875	201	206	211	216	221	226	231	236	240	245	5 2.5
876	250	255	260	265	270	275	280	285	290	295	6 3.0
877	300	305	310	315	320	325	330	335	340	345	7 3.5
878	349	354	359	364	369	374	379	384	389	394	8 4.0
879	399	404	409	414	419	424	429	433	438	443	9 4.5
880	94 448	94 453	94 458	94 463	94 468	94 473	94 478	94 483	94 488	94 493	
881	498	503	507	512	517	522	527	532	537	542	
882	547	552	557	562	567	571	576	581	586	591	
883	596	601	606	611	616	621	626	630	635	640	
884	645	650	655	660	665	670	675	680	685	689	
885	694	699	704	709	714	719	724	729	734	738	
886	743	748	753	758	763	768	773	778	783	787	
887	792	797	802	807	812	817	822	827	832	836	
888	841	846	851	856	861	866	871	876	880	885	
889	890	895	900	905	910	915	919	924	929	934	
890	94 939	94 944	94 949	94 954	94 959	94 963	94 968	94 973	94 978	94 983	4
891	988	993	998	95 002	95 007	95 012	95 017	95 022	95 027	95 032	1 0.4
892	95 036	95 041	95 046	051	056	061	066	071	075	080	2 0.8
893	085	090	095	100	105	109	114	119	124	129	3 1.2
894	134	139	143	148	153	158	163	168	173	177	4 1.6
895	182	187	192	197	202	207	211	216	221	226	5 2.0
896	231	236	240	245	250	255	260	265	270	274	6 2.4
897	279	284	289	294	299	303	308	313	318	323	7 2.8
898	328	332	337	342	347	352	357	361	366	371	8 3.2
899	376	381	386	390	395	400	405	410	415	419	9 3.6
900	95 424	95 429	95 434	95 439	95 444	95 448	95 453	95 458	95 463	95 468	
N	0	1	2	3	4	5	6	7	8	9	

Numbers 900–950 Logs 95424–97813

N	0	1	2	3	4	5	6	7	8	9
900	95 424	95 429	95 434	95 439	95 444	95 448	95 453	95 458	95 463	95 468
901	472	477	482	487	492	497	501	506	511	516
902	521	525	530	535	540	545	550	554	559	564
903	569	574	578	583	588	593	598	602	607	612
904	617	622	626	631	636	641	646	650	655	660
905	665	670	674	679	684	689	694	698	703	708
906	713	718	722	727	732	737	742	746	751	756
907	761	766	770	775	780	785	789	794	799	804
908	809	813	818	823	828	832	837	842	847	852
909	856	861	866	871	875	880	885	890	895	899
910	95 904	95 909	95 914	95 918	95 923	95 928	95 933	95 938	95 942	95 947
911	952	957	961	966	971	976	980	985	990	995
912	999	96 004	96 009	96 014	96 019	96 023	96 028	96 033	96 038	96 042
913	96 047	052	057	061	066	071	076	080	085	090
914	095	099	104	109	114	118	123	128	133	137
915	142	147	152	156	161	166	171	175	180	185
916	190	194	199	204	209	213	218	223	227	232
917	237	242	246	251	256	261	265	270	275	280
918	284	289	294	298	303	308	313	317	322	327
919	332	336	341	346	350	355	360	365	369	374
920	96 379	96 384	96 388	96 393	96 398	96 402	96 407	96 412	96 417	96 421
921	426	431	435	440	445	450	454	459	464	468
922	473	478	483	487	492	497	501	506	511	515
923	520	525	530	534	539	544	548	553	558	562
924	567	572	577	581	586	591	595	600	605	609
925	614	619	624	628	633	638	642	647	652	656
926	661	666	670	675	680	685	689	694	699	703
927	708	713	717	722	727	731	736	741	745	750
928	755	759	764	769	774	778	783	788	792	797
929	802	806	811	816	820	825	830	834	839	844
930	96 848	96 853	96 858	96 862	96 867	96 872	96 876	96 881	96 886	96 890
931	895	900	904	909	914	918	923	928	932	937
932	942	946	951	956	960	965	970	974	979	984
933	988	993	997	97 002	97 007	97 011	97 016	97 021	97 025	97 030
934	97 035	97 039	97 044	049	053	058	063	067	072	077
935	081	086	090	095	100	104	109	114	118	123
936	128	132	137	142	146	151	155	160	165	169
937	174	179	183	188	192	197	202	206	211	216
938	220	225	230	234	239	243	248	253	257	262
939	267	271	276	280	285	290	294	299	304	308
940	97 313	97 317	97 322	97 327	97 331	97 336	97 340	97 345	97 350	97 354
941	359	364	368	373	377	382	387	391	396	400
942	405	410	414	419	424	428	433	437	442	447
943	451	456	460	465	470	474	479	483	488	493
944	497	502	506	511	516	520	525	529	534	539
945	543	548	552	557	562	566	571	575	580	585
946	589	594	598	603	607	612	617	621	626	630
947	635	640	644	649	653	658	663	667	672	676
948	681	685	690	695	699	704	708	713	717	722
949	727	731	736	740	745	749	754	759	763	768
950	97 772	97 777	97 782	97 786	97 791	97 795	97 800	97 804	97 809	97 813
N	0	1	2	3	4	5	6	7	8	9

P.P.

5	
1	0.5
2	1.0
3	1.5
4	2.0
5	2.5
6	3.0
7	3.5
8	4.0
9	4.5

4	
1	0.4
2	0.8
3	1.2
4	1.6
5	2.0
6	2.4
7	2.8
8	3.2
9	3.6

Numbers 950–1000 Logs 97772–00039

N	0	1	2	3	4	5	6	7	8	9
950	97 772	97 777	97 782	97 786	97 791	97 795	97 800	97 804	97 809	97 813
951	818	823	827	832	836	841	845	850	855	859
952	864	868	873	877	882	886	891	896	900	905
953	909	914	918	923	928	932	937	941	946	950
954	955	959	964	968	973	978	982	987	991	996
955	98 000	98 005	98 009	98 014	98 019	98 023	98 028	98 032	98 037	98 041
956	046	050	055	059	064	068	073	078	082	087
957	091	096	100	105	109	114	118	123	127	132
958	137	141	146	150	155	159	164	168	173	177
959	182	186	191	195	200	204	209	214	218	223
960	98 227	98 232	98 236	98 241	98 245	98 250	98 254	98 259	98 263	98 268
961	272	277	281	286	290	295	299	304	308	313
962	318	322	327	331	336	340	345	349	354	358
963	363	367	372	376	381	385	390	394	399	403
964	408	412	417	421	426	430	435	439	444	448
965	453	457	462	466	471	475	480	484	489	493
966	498	502	507	511	516	520	525	529	534	538
967	543	547	552	556	561	565	570	574	579	583
968	588	592	597	601	605	610	614	619	623	628
969	632	637	641	646	650	655	659	664	668	673
970	98 677	98 682	98 686	98 691	98 695	98 700	98 704	98 709	98 713	98 717
971	722	726	731	735	740	744	749	753	758	762
972	767	771	776	780	784	789	793	798	802	807
973	811	816	820	825	829	834	838	843	847	851
974	856	860	865	869	874	878	883	887	892	896
975	900	905	909	914	918	923	927	932	936	941
976	945	949	954	958	963	967	972	976	981	985
977	989	994	998	99 003	99 007	99 012	99 016	99 021	99 025	99 029
978	99 034	99 038	99 043	047	052	056	061	065	069	074
979	078	083	087	092	096	100	105	109	114	118
980	99 123	99 127	99 131	99 136	99 140	99 145	99 149	99 154	99 158	99 162
981	167	171	176	180	185	189	193	198	202	207
982	211	216	220	224	229	233	238	242	247	251
983	255	260	264	269	273	277	282	286	291	295
984	300	304	308	313	317	322	326	330	335	339
985	344	348	352	357	361	366	370	374	379	383
986	388	392	396	401	405	410	414	419	423	427
987	432	436	441	445	449	454	458	463	467	471
988	476	480	484	489	493	498	502	506	511	515
989	520	524	528	533	537	542	546	550	555	559
990	99 564	99 568	99 572	99 577	99 581	99 585	99 590	99 594	99 599	99 603
991	607	612	616	621	625	629	634	638	642	647
992	651	656	660	664	669	673	677	682	686	691
993	695	699	704	708	712	717	721	726	730	734
994	739	743	747	752	756	760	765	769	774	778
995	782	787	791	795	800	804	808	813	817	822
996	826	830	835	839	843	848	852	856	861	865
997	870	874	878	883	887	891	896	900	904	909
998	913	917	922	926	930	935	939	944	948	952
999	957	961	965	970	974	978	983	987	991	996
1000	00 000	00 004	00 009	00 013	00 017	00 022	00 026	00 030	00 035	00 039
N	0	1	2	3	4	5	6	7	8	9

P.P.

	5
1	0.5
2	1.0
3	1.5
4	2.0
5	2.5
6	3.0
7	3.5
8	4.0
9	4.5

	4
1	0.4
2	0.8
3	1.2
4	1.6
5	2.0
6	2.4
7	2.8
8	3.2
9	3.6

Numbers 1000–1050 Logs 0000000–0215614

N	0	1	2	3	4	5	6	7	8	9
1000	000 0000	0434	0869	1303	1737	2171	2605	3039	3473	3907
01	4341	4775	5208	5642	6076	6510	6943	7377	7810	8244
02	8677	9111	9544	9977	*0411	*0844	*1277	*1710	*2143	*2576
03	001 3009	3442	3875	4308	4741	5174	5607	6039	6472	6905
04	7337	7770	8202	8635	9067	9499	9932	*0364	*0796	*1228
05	002 1661	2093	2525	2957	3389	3821	4253	4685	5116	5548
06	5980	6411	6843	7275	7706	8138	8569	9001	9432	9863
07	003 0295	0726	1157	1588	2019	2451	2882	3313	3744	4174
08	4605	5036	5467	5898	6328	6759	7190	7620	8051	8481
09	8912	9342	9772	*0203	*0633	*1063	*1493	*1924	*2354	*2784
1010	004 3214	3644	4074	4504	4933	5363	5793	6223	6652	7082
11	7512	7941	8371	8800	9229	9659	*0088	*0517	*0947	*1376
12	005 1805	2234	2663	3092	3521	3950	4379	4808	5237	5666
13	6094	6523	6952	7380	7809	8238	8666	9094	9523	9951
14	006 0380	0808	1236	1664	2092	2521	2949	3377	3805	4233
15	4660	5088	5516	5944	6372	6799	7227	7655	8082	8510
16	8937	9365	9792	*0219	*0647	*1074	*1501	*1928	*2355	*2782
17	007 3210	3637	4064	4490	4917	5344	5771	6198	6624	7051
18	7478	7904	8331	8757	9184	9610	*0037	*0463	*0889	*1316
19	008 1742	2168	2594	3020	3446	3872	4298	4724	5150	5576
1020	6002	6427	6853	7279	7704	8130	8556	8981	9407	9832
21	009 0257	0683	1108	1533	1959	2384	2809	3234	3659	4084
22	4509	4934	5359	5784	6208	6633	7058	7483	7907	8332
23	8756	9181	9605	*0030	*0454	*0878	*1303	*1727	*2151	*2575
24	010 3000	3424	3848	4272	4696	5120	5544	5967	6391	6815
25	7239	7662	8086	8510	8933	9357	9780	*0204	*0627	*1050
26	011 1474	1897	2320	2743	3166	3590	4013	4436	4859	5282
27	5704	6127	6550	6973	7396	7818	8241	8664	9086	9509
28	9931	*0354	*0776	*1198	*1621	*2043	*2465	*2887	*3310	*3732
29	012 4154	4576	4998	5420	5842	6264	6685	7107	7529	7951
1030	8372	8794	9215	9637	*0059	*0480	*0901	*1323	*1744	*2165
31	013 2587	3008	3429	3850	4271	4692	5113	5534	5955	6376
32	6797	7218	7639	8059	8480	8901	9321	9742	*0162	*0583
33	014 1003	1424	1844	2264	2685	3105	3525	3945	4365	4785
34	5205	5625	6045	6465	6885	7305	7725	8144	8564	8984
35	9403	9823	*0243	*0662	*1082	*1501	*1920	*2340	*2759	*3178
36	015 3598	4017	4436	4855	5274	5693	6112	6531	6950	7369
37	7788	8206	8625	9044	9462	9881	*0300	*0718	*1137	*1555
38	016 1974	2392	2810	3229	3647	4065	4483	4901	5319	5737
39	6155	6573	6991	7409	7827	8245	8663	9080	9498	9916
1040	017 0333	0751	1168	1586	2003	2421	2838	3256	3673	4090
41	4507	4924	5342	5759	6176	6593	7010	7427	7844	8260
42	8677	9094	9511	9927	*0344	*0761	*1177	*1594	*2010	*2427
43	018 2843	3259	3676	4092	4508	4925	5341	5757	6173	6589
44	7005	7421	7837	8253	8669	9084	9500	9916	*0332	*0747
45	019 1163	1578	1994	2410	2825	3240	3656	4071	4486	4902
46	5317	5732	6147	6562	6977	7392	7807	8222	8637	9052
47	9467	9882	*0296	*0711	*1126	*1540	*1955	*2369	*2784	*3198
48	020 3613	4027	4442	4856	5270	5684	6099	6513	6927	7341
49	7755	8169	8583	8997	9411	9824	*0238	*0652	*1066	*1479
1050	021 1893	2307	2720	3134	3547	3961	4374	4787	5201	5614
N	**0**	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**

Numbers 1050–1100 Logs 0211893–0417479

N	0	1	2	3	4	5	6	7	8	9
1050	021 1893	2307	2720	3134	3547	3961	4374	4787	5201	5614
51	6027	6440	6854	7267	7680	8093	8506	8919	9332	9745
52	022 0157	0570	0983	1396	1808	2221	2634	3046	3459	3871
53	4284	4696	5109	5521	5933	6345	6758	7170	7582	7994
54	8406	8818	9230	9642	*0054	*0466	*0878	*1289	*1701	*2113
55	023 2525	2936	3348	3759	4171	4582	4994	5405	5817	6228
56	6639	7050	7462	7873	8284	8695	9106	9517	9928	*0339
57	024 0750	1161	1572	1982	2393	2804	3214	3625	4036	4446
58	4857	5267	5678	6088	6498	6909	7319	7729	8139	8549
59	8960	9370	9780	*0190	*0600	*1010	*1419	*1829	*2239	*2649
1060	025 3059	3468	3878	4288	4697	5107	5516	5926	6335	6744
61	7154	7563	7972	8382	8791	9200	9609	*0018	*0427	*0836
62	026 1245	1654	2063	2472	2881	3289	3698	4107	4515	4924
63	5333	5741	6150	6558	6967	7375	7783	8192	8600	9008
64	9416	9824	*0233	*0641	*1049	*1457	*1865	*2273	*2680	*3088
65	027 3496	3904	4312	4719	5127	5535	5942	6350	6757	7165
66	7572	7979	8387	8794	9201	9609	*0016	*0423	*0830	*1237
67	028 1644	2051	2458	2865	3272	3679	4086	4492	4899	5306
68	5713	6119	6526	6932	7339	7745	8152	8558	8964	9371
69	9777	*0183	*0590	*0996	*1402	*1808	*2214	*2620	*3026	*3432
1070	029 3838	4244	4649	5055	5461	5867	6272	6678	7084	7489
71	7895	8300	8706	9111	9516	9922	*0327	*0732	*1138	*1543
72	030 1948	2353	2758	3163	3568	3973	4378	4783	5188	5592
73	5997	6402	6807	7211	7616	8020	8425	8830	9234	9638
74	031 0043	0447	0851	1256	1660	2064	2468	2872	3277	3681
75	4085	4489	4893	5296	5700	6104	6508	6912	7315	7719
76	8123	8526	8930	9333	9737	*0140	*0544	*0947	*1350	*1754
77	032 2157	2560	2963	3367	3770	4173	4576	4979	5382	5785
78	6188	6590	6993	7396	7799	8201	8604	9007	9409	9812
79	033 0214	0617	1019	1422	1824	2226	2629	3031	3433	3835
1080	4238	4640	5042	5444	5846	6248	6650	7052	7453	7855
81	8257	8659	9060	9462	9864	*0265	*0667	*1068	*1470	*1871
82	034 2273	2674	3075	3477	3878	4279	4680	5081	5482	5884
83	6285	6686	7087	7487	7888	8289	8690	9091	9491	9892
84	035 0293	0693	1094	1495	1895	2296	2696	3096	3497	3897
85	4297	4698	5098	5498	5898	6298	6698	7098	7498	7898
86	8298	8698	9098	9498	9898	*0297	*0697	*1097	*1496	*1896
87	036 2295	2695	3094	3494	3893	4293	4692	5091	5491	5890
88	6289	6688	7087	7486	7885	8284	8683	9082	9481	9880
89	037 0279	0678	1076	1475	1874	2272	2671	3070	3468	3867
1090	4265	4663	5062	5460	5858	6257	6655	7053	7451	7849
91	8248	8646	9044	9442	9839	*0237	*0635	*1033	*1431	*1829
92	038 2226	2624	3022	3419	3817	4214	4612	5009	5407	5804
93	6202	6599	6996	7393	7791	8188	8585	8982	9379	9776
94	039 0173	0570	0967	1364	1761	2158	2554	2951	3348	3745
95	4141	4538	4934	5331	5727	6124	6520	6917	7313	7709
96	8106	8502	8898	9294	9690	*0086	*0482	*0878	*1274	*1670
97	040 2066	2462	2858	3254	3650	4045	4441	4837	5232	5628
98	6023	6419	6814	7210	7605	8001	8396	8791	9187	9582
99	9977	*0372	*0767	*1162	*1557	*1952	*2347	*2742	*3137	*3532
1100	041 3927	4322	4716	5111	5506	5900	6295	6690	7084	7479
N	0	1	2	3	4	5	6	7	8	9

APPENDIX D

Selected Reference List

Barlow's Tables of Squares, Cubes, Square Roots, Cube Roots, and Reciprocals of all integer numbers up to 10,000. E. and F. N. Spon, Ltd., London.

Chapin, F. S., *Field Work and Social Research,* The Century Co., New York.

Chaddock, R. E., *Principles and Methods of Statistics,* Houghton Mifflin Co., Boston.

Dubois, Florence, *A Guide to Statistics of Social Welfare in New York City,* Welfare Council of New York City, New York.

Dunlap, J. W., and Kurtz, A. K., *Handbook of Statistical Monographs, Tables and Formulas,* World Book Co., Yonkers-on-Hudson, New York.

Ezekiel, Mordicai, *Methods of Correlation Analysis,* John Wiley & Sons, New York.

Fry, C. Luther, "Making Use of Census Data," *Jour. Amer. Stat. Ass'n,* Columbia University, New York, June, 1930.

Glover, J. W., *Tables of Applied Mathematics in Finance, Insurance and Statistics,* Millard Press, Ann Arbor, Mich.

Hexter, Maurice B., *Social Consequences of Business Cycles,* Houghton Mifflin Co., Boston.

Journal of the American Statistical Association, Columbia University, New York.

Kelley, T. L., *Statistical Method,* Macmillan Co., New York.

King, W. I., *Index Numbers Elucidated,* Longmans, Green & Co., New York.

Macaulay, F. R., *The Smoothing of Time Series,* National Bureau of Economic Research, Inc., New York.

McMillen, A. W., *Measurement in Social Work,* University of Chicago Press, Chicago.

Mills, F. C., *Statistical Methods,* Henry Holt & Co., New York.

Mudgett, Bruce D., *Statistical Tables and Graphs*, Houghton Mifflin Co., Boston.

Pearl, Raymond, *Medical Biometry and Statistics*, W. B. Saunders Co., Philadelphia.

Pearson, Karl, *Tables for Statisticians and Biometricians*, Cambridge University Press, Cambridge.

Proceedings of the American Statistical Association, Columbia University, New York.

Rice, Stuart A. (editor), *Statistics in Social Studies*, University of Pennsylvania Press, Philadelphia.

Rietz, H. L. (editor), *Handbook of Mathematical Statistics*, Houghton Mifflin Co., Boston.

Schemeckebier, L. F., *The Statistical Work of the National Government*, Johns Hopkins Press, Baltimore.

Thomas, Dorothy S., *Social Aspects of the Business Cycle*, Routledge, London.

Thurstone, L. L., *The Fundamentals of Statistics*, Macmillan Co., New York.

Thurstone, L. L., and Chave, E. J., *The Measurement of Attitude*, University of Chicago Press, Chicago.

Walker, Helen D., *Studies in the History of Statistical Method*, Williams and Wilkins, Baltimore.

Weld, L. D., *Theory of Errors and Least Squares*, Macmillan Co., New York.

INDEX